Challenges in analysis and processing of spontaneous speech

CAPSS

Edited by
Mária Gósy and Tekla Etelka Gráczi

RESEARCH INSTITUTE FOR LINGUISTICS
HUNGARIAN ACADEMY OF SCIENCES

CHALLENGES IN ANALYSIS AND PROCESSING OF SPONTANEOUS SPEECH

Edited by
Mária GÓSY and Tekla Etelka GRÁCZI

CHALLENGES IN ANALYSIS AND PROCESSING OF SPONTANEOUS SPEECH

Selected and peer-reviewed papers of the workshop entitled
Challenges in Analysis and Processing of Spontaneous Speech (Budapest, 2017)

CHALLENGES IN ANALYSIS AND PROCESSING OF SPONTANEOUS SPEECH

Edited by
Mária GÓSY and Tekla Etelka GRÁCZI

RESEARCH INSTITUTE FOR LINGUISTICS
HUNGARIAN ACADEMY OF SCIENCES

Budapest, 2018

The book consists of selected and peer-reviewed papers of the workshop entitled *Challenges in Analysis and Processing of Spontaneous Speech* (Budapest, 2017)

Publishing was funded by the National OTKA 108762 Project.

ISBN 978-963-9074-75-0 (printed edition)

ISBN 978-963-9074-76-7 (online edition)

DOI: http://doi.org/10.18135/CAPSS (online edition)

Technical editing: Tekla Etelka GRÁCZI

Cover: Dorottya GYARMATHY

Publisher:

RESEARCH INSTITUTE FOR LINGUISTICS
HUNGARIAN ACADEMY OF SCIENCES

Prof. Gábor PRÓSZÉKY, director of the Research Institute for Linguistics of the Hungarian Academy of Sciences

Printed in Hungary: OOK Press, Veszprém

Budapest, 2018

CONTENTS

PREFACE

In view of the rapid growth of various aspects in speech research, this edited volume addresses the issue of spontaneous speech research.

The problem area of spontaneous speech raises various questions for all fields of the speech sciences. Studies of the manifold challenges of its analysis and processing are introduced in this book: The topics are about how a speaker makes use of speech organs to deliver thoughts from larynx movements through co-articulation processes to demands of various communication settings. The book is also about atypical speech, disfluencies and speech synthesis, and it includes a speaker age estimation research. Acoustic-phonetic analyses deal with irregular phonation, glottal markings, and phrase-final lengthening. The chapters show a wide range of topics that researchers focus on worldwide. The cohesion of the chapters lies in their diversity with a common approach to discuss processes, phenomena, and results of some kind in speech science.

The themes of the papers represent shared knowledge of the research community that has accumulated during the past decades. The majority of the papers are devoted to specific questions in diverse languages (e.g., English, Georgian, and Hungarian). There is a major emphasis on spontaneous speech, and many topics confirm evidence that spontaneously produced speech is very complex, and speakers show large individual differences (as expected). Some other papers investigate read speech to better understand how the specific issue can be extended later to spontaneous speech.

The ten papers of this volume were selected from the talks of an international workshop that was organized last year in Budapest and was entitled Challenges in analysis and processing of spontaneous speech (CAPSS). The papers underwent the usual peer-reviewing process to be included into this book. The workshop provided a unique exchange forum for researchers working on all kinds of research fields focusing on relevant questions of spontaneous speech. Our intention in editing this book was that the selected papers should reflect the overall research interest of the participants of the workshop, on the one hand, and ongoing phonetic investigations in these days, on the other.

The book is dedicated to all those who are interested in the analysis of spontaneous speech from various aspects and in diverse languages. As editors, we do hope that the studies of the book and their results will encourage our

fellow researchers to discuss these issues further, and make our common knowledge take a large step forward in this cognitive process.

Finally, we would like to thank our authors for having worked so hard to turn their talks into high-level scientific papers and for being enthusiastic in improving their manuscripts in response to the reviewers' comments. We would also like to thank our reviewers who accepted our invitation to review the papers, and acted quickly both when they received the original manuscripts and when they re-read the revised versions.

We hope to provide you with a bookful of new thought-provoking results on interesting topics to initiate a growing interest in spontaneous speech research.

The editors
Mária Gósy and *Tekla Etelka Gráczi*

Budapest, June 2018.

INVESTIGATING CLEAR SPEECH ADAPTATIONS IN SPONTANEOUS SPEECH PRODUCED IN COMMUNICATIVE SETTINGS

Outi TUOMAINEN & Valerie HAZAN
Department of Speech Hearing and Phonetic Sciences,
University College London (UCL), London, UK.
o.tuomainen@ucl.ac.uk, v.hazan@ucl.ac.uk

Abstract
In order to investigate the clear speech adaptations that individuals make when communicating in intelligibility-challenging conditions, it would seem essential to examine speech that is produced in interaction with a conversational partner. However, much of the literature on clear speech adaptations has been based on the analysis of sentences that talkers were instructed to read clearly. In this chapter, we review methods for eliciting spontaneous speech in interaction for the purpose of investigating clear speech phenomena. We describe in more detail the Diapix task (Van Engen et al., 2010) and DiapixUK picture pairs (Baker & Hazan, 2011) which have been used in the production of large corpora investigating clear speech adaptations. We present an overview of the analysis of spontaneous speech and clear speech adaptations from the LUCID corpora that include spontaneous speech recordings from children, young and older adults.

Keywords: spontaneous speech, clear speech, speech corpus

1 Introduction

The aim of our research is to investigate the acoustic-phonetic adaptations that individuals make to their speech to be able to communicate effectively in challenging environments. These are most often described as talking in a 'clear speaking style' or with 'clear speech' and have been the subject of many investigations over the last 30 or so years. This chapter will review the experimental approaches that have typically been used to investigate clear speech and will argue for investigating this phenomenon using spontaneous speech produced in interaction during a communicative task. We describe techniques for recording speech corpora that have taken this approach. Finally, we present some key findings from three linked studies that have investigated clear speech adaptations in children, young and older adults.

Most speech communication occurs in situations that are less than ideal. These situations have been referred to in the literature as 'adverse' or 'challenging' conditions but in fact represent many of the conditions that we encounter in

DOI: http://doi.org/10.18135/CAPSS.9

our everyday lives. For example, a study that investigated the typical listening environments of a small group of older adults, using an experiential sampling method, found that for 40% of the reported time, individuals were in an environment that was 'a bit noisy' and that about 10% of the time was in situations that were noisy or very noisy (Hasan et al., 2014). A recent review paper (Mattys et al., 2012) provided a useful classification of causes of adverse conditions for speech communication. First, there can be environmental/transmission degradation such as the presence of noise or other voices in the environment, or a high degree of reverberation. There can be receiver limitations, such as the presence of a hearing loss, the lack of shared language knowledge or lack of available resources due to a high degree of cognitive load. To this classification can be added speaker limitations, such as the presence of a language impairment or speaking in an unknown accent. All these situations can lead to communication difficulties that can result in disfluencies, multiple requests for repetitions or clarifications, and misunderstandings needing repair.

Young adults are skilled at making adaptations to their conversational speaking style in order to overcome the effects of these types of degradation. The adjustments that they make typically require greater effort on the part of the speaker but are evidence, as suggested by Lindblom in his Hyper-Hypo model of speech production (Lindblom, 1990), that speech produced in interaction is listener-oriented and for the benefit of efficient communication. While talkers typically aim to minimise the degree of articulatory effort that they expend in their conversational speaking style, this will be increased for the sake of efficient communication if the type of communication barriers mentioned above are affecting intelligibility. Talkers continuously assess the level of understanding of their interlocutor via the appropriateness of their responses, the frequency of requests for clarification, pauses, and hesitations. In conditions in which individuals are conversing in a very noisy room, for example, they may adopt a clear speaking style, but if communication is progressing well, talkers might start to reduce the effort that they are making to speak more clearly. However, if there is a breakdown and need for repair, the degree of clear speaking style is likely to increase. This type of speaking style is therefore highly dynamic and listener-focused.

The adaptations that are made in clear speech can be at the level of acoustic-phonetic or linguistic adjustments. Excellent reviews of the types of adaptations made in clear speech can be found in Smiljanic and Bradlow (2009), Mattys et al. (2012) and Cooke et al. (2014). In summary, acoustic-phonetic adaptations can include reductions in articulation rate, increases in pause frequency and in fundamental frequency, shifts in the energy distribution in the voice and vowel hyper-articulation. Linguistic adaptations can entail the use of more frequent words, reductions in sentence length and complexity or changes in lexical

diversity (e.g., Granlund et al., 2018). Adaptations are, to a degree, tailored to best counter the interference that the interlocutor may be experiencing (e.g., Cooke & Lu, 2010; Hazan & Baker, 2011). In summary, speaking style adaptations are a fairly skilled aspect of speech production and are often essential for efficient and effective communication in challenging situations.

Given that clear speaking styles are likely to be strongly dependent on the interaction between talker and listener, and on the degree of difficulty experienced by one or both conversational partners, it would appear of paramount importance to involve interaction and communicative intent in the study of clear speech adaptations. However, the great majority of studies of clear speaking styles which led to the findings listed above have involved an approach where communicative intent was absent. Indeed, a typical approach has been to instruct the participant to read a set of sentences 'normally' or using a conversational style, and then to ask participants to read the sentences again, but this time as if speaking to a person who is hearing impaired or who is a non-native speaker. In terms of experimental control, this approach has a number of advantages over spontaneous speech in that the speech to be analysed is consistent across talkers and across speaking styles. This is a great advantage when measuring the acoustic characteristics of speech which can be affected by coarticulation, lexical content and many other sources of variation. However, such read speech has a number of shortcomings. First, the recorded speech materials lack communicative intent and the dynamic adjustments that talkers make to their speech in natural interactions. Also, in studies involving read materials, the participant is merely a 'talker' whereas true communication involves a participant as both listener and talker, and, more often than not, doing another task while communicating. The added cognitive load involved in such interactions could affect aspects of speech production (e.g., Nip & Green, 2013).

Another important shortcoming of using read speech for investigations into clear speech adaptations is that there is evidence that the clear speech that is recorded when participants are instructed to read clearly differs in some respects from naturally-elicited clear speech. For example, Hazan and Baker (2011) found that, for a same set of talkers, clear speech that was elicited via instruction in read sentences showed more extreme changes in at least certain acoustic-phonetic characteristics than spontaneous speech produced to counteract intelligibility-challenging conditions. In a study comparing different types of instructions to speak clearly with speech directed at an interlocutor, Scarborough and Zellou (2013) found that instructions to speak clearly led to greater changes in vowel duration and more greatly-hyperarticulated vowels than the naturally-elicited clear speech. In the same study, when speech samples were presented in a lexical decision task, listeners responded more quickly to the naturally-elicited clear speech than to the speech spoken 'as if to someone who is hard of hearing'.

Differences were also found between 'real' and 'imagined' foreigner-directed clear speech and authors argued for the use of 'communicatively authentic elicitation tasks' in studies of clear speaking styles (Scarborough et al. 2007). Finally, within 'instructed' clear speech, the perceptual benefit of clear speaking styles varies with the type of instruction given (Lam & Tjaden, 2013).

Recently, there has been a move towards investigating clear speech adaptations in corpora of spontaneous speech collected while pairs of participants were involved in a problem-solving task (for a review, see Cooke et al., 2014). These dialogues may still be far from natural communication, as they are recorded in laboratory conditions and involved talkers carrying out a specific problem-solving task in order to maintain some control over the content and duration of the interaction. However, they provide an important half-way house between read speech and totally unstructured spontaneous speech. Another advantage of this approach is that they model the kind of multi-tasking and sharing of cognitive resources that occurs in much natural communication. Using this type of interactive task, clear speech adaptations can be naturally elicited by, for example, adding noise in the background while the task is being carried out, and such speech is then compared to speech recorded when there was no interference affecting participants.

An early example of a collaborative problem-solving task used in the recording of speech corpora is the 'Map Task' which was used in the development of the HCRC Map Task corpus (Anderson et al., 1991). The Map Task involves 'instruction givers' having to communicate details of a map route and of different key elements on the map to 'instruction followers' who have no indication of the route on their map; the two maps can also differ in some key elements. Task success can be measured using a deviation score from the accurate route. The task has been used in a number of studies investigating, for example, word segmentation cues (White et al., 2010) and the role of visual cues in communicating information (Anderson et al., 1991). The speech recorded using this task includes many direction-giving commands and requests. However, the maps are highly simplified and accompanied by labels, so there is little variation in the lexical content produced.

Another approach to elicit spontaneous speech dialogues has been to use popular problem-solving puzzles. In Cooke and Lu (2010), pairs of participants cooperatively completed Sudoku puzzles, which provided many repetitions of number words. Crosswords have also been used (Crawford et al., 1994); these can lead to the use of a wider range of lexical items than Sudoku. However, in both these approaches, both participants see the same information and one participant can dominate the task without requiring much input from the other. It is also the case that individuals vary widely in their skill and interest in completing word- or number-based puzzles and such tasks may be too complex

for very young or much older participants. Tangram puzzles (Clark & Wilkes-Gibbs, 1986) have also been used in many studies of speech in interaction. Tangrams are visual puzzles in which sets of shapes (squares, triangles) have to be assembled in a specific way to form a shape. Tangrams have the advantage of involving less skill than word or number puzzles and can also be used in many different permutations, so are less limited than for example Map Tasks which have to be carefully constructed. The type of interactions they elicit is still fairly limited though and would include a high proportion of short commands. Tangrams have been used in studies such as Murfitt and McAllister (2001).

A task that was recently developed for investigations of clear speech adaptations in children is the Grid Task (Granlund et al., 2018). In this task, each participant is given a grid with pictures, an empty grid with squares with coloured numbers and a tray containing five different drawn versions of 16 keywords that formed minimal pairs (e.g., *peach* - *beach*). The aim of the task was for each conversational partner, without being able to see each other's grids, to replicate their partner's grid in their empty grid. In order to do this, they had to converse with their conversational partner to find for each of the 16 boxes on the grid the correct keyword, the correct version of the keyword, and the correct location of the keyword (see Granlund et al., 2018, for an example of the grid and picture materials). This task very much engaged the children and the need to differentiate five different representations of an object led to variation in the lexical content. Multiple iterations of each keyword were produced as well as many repetitions of numbers and colours.

Finally, some spontaneous speech corpora have taken the approach of recording conversations between two interlocutors on everyday topics that are likely to elicit a range of different views and opinions. For example, in the BEA corpus (Gósy, 2012), the conversation module involved the participant, interviewer and a third person discussing topics such as 'marriage vs cohabitation' or 'secondary school final exams'. This approach is likely to elicit more natural spontaneous speech than problem-based tasks, although some individuals might be more reluctant to express personal views and therefore produce less speech.

2 The Diapix task

A recent task, which is becoming widely used for recordings of speech in interaction, is the Diapix task (Van Engen et al., 2010), which was first developed to compare conversational speech interactions between pairs of native and non-native speakers. Diapix involves pairs of participants engaged in a 'spot the difference' picture task. Each participant is presented with a different version of the same cartoon-style picture, and both have to collaborate to find the differences between the two pictures without seeing each other's picture. A set

of 12 carefully-designed picture pairs developed by Baker and Hazan, the DiapixUK pictures, have been used in a number of different studies and are available as supplementary materials in Baker and Hazan (2011).

The design of the DiapixUK picture pairs was done with great care (see Baker & Hazan, 2011). There are four picture-pairs for each of three main scenes: beach, street and farm, each containing 12 differences. A number of factors were controlled in these pictures, such as the position of differences within the picture, the type of difference to be found (presence/absence of object vs change in object), which of the two pictures was key to finding the difference (if absence of an object) in order to ensure that both participants had to take an active part in the task. The pictures were also made to be quite humorous to maintain interest and encourage more relaxed conversation. Analyses in Baker and Hazan (2011) revealed that, unless one talker was instructed to take the lead, Diapix led to balanced speech being recorded for both participants (Talker A: 51%, Talker B 49%) which differed from the Map Task where the instruction giver contributed 68% of words. Also, after participants had completed a practice picture, there was no learning effect when several Diapix tasks were run in succession, as shown by task transaction time. The level of difficulty of the pictures was also found to be consistent, as shown by non-significant differences in task transaction time, although there was greater variation for the later-developed pictures (named beach 4, street 4, farm 4).

The DiapixUK picture pairs (Baker & Hazan, 2011) were developed with reusability in mind: they were designed using Adobe Photoshop software with each object placed on a different layer so that the pictures could easily be edited if further changes were needed. For example, these pictures have been adapted for use with Finnish (Granlund et al., 2012), Spanish (Lecumberri et al., 2017) and Swedish participants (Sørensen et al., 2017) by changing some of the written elements (such as shop names) in the pictures. The DiapixUK picture pairs have also been adapted to investigate regional dialectal differences in British Sign Language (Stamp et al., 2016). It should be noted that certain visual elements, such as the fact that men were wearing socks with their sandals on the beach, are rather culturally-biased, but these objects can be edited or removed. Unlike word- or number-based puzzles that have different demands across groups varying in age and ability, Diapix is well suited for a wide variety of participants, including clinical populations, as the task can be solved using simple vocabulary and grammar. The DiapixUK picture pairs have been used, without alterations, with participants aged 8 to 85 years.

As with the Map Task, the differences in Diapix were designed to encourage the repetition of specific keywords, in this case words from /p/-/b/ and /s/-/ʃ/ minimal pairs, so that segmental contrasts could be analysed. It is also possible to obtain measures of vowel space by accumulating vowel formant measures for

point vowels that occur frequently in content words throughout the spontaneous speech interactions (e.g., Pettinato et al., 2016). However, rather than detailed segmental analyses, Diapix recordings are more commonly used to investigate more global characteristics of speech, such as measures of articulation rate, fundamental frequency and long-term average spectrum. They can also be used to obtain measures related to the conversational interaction, such as rate and type of disfluencies or of repairs.

Another type of measure that can be obtained from Diapix is a measure of communication 'success' or efficiency. Indeed, difficulties in communication result in increased pausing, disfluencies, repetitions, elaborations that all lengthen the time needed to complete the task successfully. Measures of communication efficiency include task transaction time (e.g., Van Engen et al., 2010), the number of differences found in a set time, and the frequency of communication breakdowns (McInerney & Walden, 2013). Another measure which can be of use in evaluating participant dominance is the time spent by each holding the floor in the interactions while they are completing the task (Sørensen et al., 2017).

In order to use the Diapix task to naturally elicit clear speech adaptations, it is necessary to make communication difficult for one or both participants in the interaction. One means of achieving this while also maintaining high quality and 'clean' recordings for each of the participants in the interaction involves seating participants in separate sound-treated booths and having them communicate via headsets. Recordings are controlled via a computer in a separate control room. One or both audio channels can be manipulated to degrade the signal being transmitted from one participant to the other in real time. This can be done using a vocoder, for example, by adding noise or babble to the channel or by using software such as HELPS (Zurek & Desloge, 2007) to simulate a sensorineural hearing loss. The aim is to naturally elicit clear speech adaptations in the 'unimpaired' participant who has to make him or herself clear for their conversational partner who has difficulty hearing them. The advantage of using this approach is that the degree of degradation can be carefully controlled, with no headphone 'leakage' heard by the other participant, and also that the speech of each participant is recorded on a separate channel with no audible interference. This is particularly important if the speech is to be used for acoustic analyses. A simpler recording set-up with both participants seated in the same room and recorded on a single channel can work well if the aim is to collect more general measures of task duration or measures of disfluencies, for example, and if recording quality is less paramount.

To date, Diapix has been used in studies of clear speech adaptations in young adults (Hazan & Baker, 2011), children with typical hearing (Hazan et al., 2016) and hearing loss (Granlund et al., 2018), older adults (Tuomainen & Hazan,

2016), native and nonnative talkers (Van Engen et al., 2010). It has also been used to examine how speech characteristics vary when speaking in one's first and second language (Lecumberri et al., 2017) and to investigate phenomena of talker convergence (Kim et al., 2011; Solanski et al., 2015; Stamp et al., 2016). Recently, the Spanish version of the Diapix pictures was used for sociolinguistic purposes, in a language contact study in Colombia. The value of the Diapix task for investigating speech interactions in more clinical settings is also being recognised. A pilot study (McInerney & Walden, 2013) used Diapix interactions to evaluate the effect of assistive listening devices (ALD) on communication efficiency in older adults with hearing loss, using the frequency of communication breakdowns as efficiency measure.

3 LUCID corpora

Three major corpora were collected at UCL consecutively over a nine year period using the Diapix task and DiapixUK picture sets. The first corpus (LUCID: London UCL Clear speech in Interaction Database) includes extensive speech recordings for 40 native Southern British English adults (20 female) aged between 18 and 29 years old (mean age: 23 years). Participants were monolingual and had normal hearing thresholds. They were recorded in a number of easy and challenging communicative conditions, with three Diapix picture tasks per condition. In the easy NORM condition, both participants could hear each other without interference. Challenging conditions included the vocoder condition (VOC: talker B heard talker A via a three-channel noise-excited vocoder) which was done by all participants. For this condition and the NORM condition, the conversational partners were known to each other. There were two further conditions, each carried out by half of the participants: in the Babble condition (BAB) talker B heard talker A's voice mixed with 8-talker babble at approximately 0 dB SNR and in the L2 condition, talker B was a low-proficiency L2 speaker. For these two conditions, Talker B was a confederate not known to the key participant. Further details about the design of the challenging conditions and of the resulting corpus can be found in Hazan and Baker (2011). The LUCID corpus also included read sentences and picture naming in two speaking styles. In the casual style, participants were instructed to read 'casually as if talking to a friend' and in the clear speaking style, they were instructed to read 'clearly as if talking to someone who is hearing impaired'. The corpus includes stereo audio files for two-way dialog (wav format) and individual wav files for each speaker as well as word-aligned orthographic transcriptions (in Praat TextGrid format). Note that annotations are not available for the speech produced by the L2 confederates. This corpus is available online (following password request) and stored within the OSCAAR archive based at Northwestern University (https://oscaar.ci.northwestern.edu/).

The kidLUCID corpus includes Diapix recordings from 96 children and adolescents aged between 9 and 14 years inclusive (50 F, 46 M, mean: 11;8 years). Participants were non-bilingual native Southern British English speakers who reported no history of hearing or language impairments. In this corpus, only one Diapix task was carried out per condition. Diapix was carried out in three of the conditions also included in the LUCID corpus for comparability: the NORM, BAB and VOC conditions. Further details about corpus design are available in Hazan et al. (2016). This corpus, together with word-level annotations in Praat Textgrids, is also available within the OSCAAR archive.

The most-recently collected elderLUCID corpus includes speech from 83 single-sex pairs of native Southern British English adult talkers between the ages of 19 and 84 years. Talker A participants were from two distinct age groups: 'younger adults' (YA) aged 19-26 years (15 F, 11 M; Mean: 21.5 yrs) and 'older adults' (OA) aged 65-84 years (30 F, 27 M, Mean: 72.5 yrs). Participants in Talker B role were always younger adults (N = 83, between 18-30 years of age) of the same sex as the Talker A. Participants reported no history of speech or language impairments. YA participants all had normal hearing thresholds. OA participants had either normal hearing (OANH: 14 F, 13M), i.e., hearing threshold of < 20 dB between 250-4000 Hz, or a mild hearing loss (OAHL: 16 F, 14 M), with hearing threshold of < 45 dB between 250-4000 Hz, typical of early stages of age-related hearing loss or presbycusis. In addition to the normal (NORM) condition, there were three challenging conditions carried out by all participants. In the hearing loss simulation condition (HLS), the voice of Talker A was processed in real time through the HELPS software (Zurek & Desloge, 2007) mimicking the effect of severe-to-profound age-related hearing loss before being transmitted to Talker B. In the BABBLE (BAB-1) condition, the speech of talker A was mixed with the same 8-talker babble as used in the previous LUCID corpora before being channelled through to the confederate's headphones, at a difficulty level equated to the HLS conditions via a Modified Rhyme Task (MRT). In the other BABBLE (BAB-2) condition, both talkers heard the same babble as in BAB-1 but at 0 dB SNR. One Diapix task was carried out per condition. For the same participants, further speech is available for the same conditions using a sentence repetition task where participants had to read sentences to Talker B who had to repeat them back. The elderLUCID corpus will be made available on request from the authors.

As the three corpora were collected in separate studies, they are not fully comparable in terms of the methodology used in their collection. This reflects the difficult decision to be made between maintaining full compatibility across related corpora collected over a period of years, and making necessary improvements or adjustments, or simply practical changes to aid recruitment. For example, in the LUCID and kidLUCID corpora, participants carried out the

Diapix task with a friend (for the NORM and VOC conditions only in LUCID) whereas in the elderLUCID corpus, participants were paired with a young adult conversational partner they had just met. This difference in degree of familiarity is likely to have an effect in the level of alignment between speakers during their interactions, for example.

4 Summary of findings on spontaneous speech across the lifespan

In this section, we summarise the main findings resulting from the analyses of the suprasegmental features of articulation rate, fundamental frequency and long-term spectrum characteristics in the three LUCID corpora. A description of the post-processing stages used in analysing these acoustic features is in Hazan et al. (2016).

First, the availability of spontaneous speech data collected using a common task in related studies with children aged 9 to 14, young adults and older adults aged 65 to 85 enables us to examine age trends for conversational speech produced in good communicative conditions but without face-to-face visual cues. Trends for articulation rate (syllables produced per second) in conversational speech showed an inverted U shape with children up to the age of 11 speaking at a slower speech rate than young adults (Hazan et al., 2016), but older adults in the 65-85 year age range also speaking at a lower articulation rate than young adults (Tuomainen & Hazan, 2016a), see Figure 1. This is consistent with the findings of Jacewicz et al. (2010), for example, for spontaneous speech monologues and Bóna (2014) for a variety of speaking styles.

Fundamental frequency measures were calculated using the de Looze and Hirst formula described in Hazan et al. (2016). Changes in mean fundamental frequency (see Figure 2) followed trends in terms of talker sex and age that were expected from the literature with differentiation on the basis of talker sex appearing around the age of 13-14 years followed by a steep reduction in fundamental frequency for male speakers in young adulthood and a more gradual reduction for female speakers (Hazan, 2017). Some studies have identified increases in mean fundamental frequency in older males and decreases in post-menopausal female talkers but, although such a trend was present, this effect was not statistically significant in our elderLUCID corpus.

Normalized pitch range (75th -25th percentile range of fundamental frequency values calculated over the aggregated speech per condition and converted to semitones relative to 1 Hz) also showed a U-shape but in the opposite direction than articulation rate: both 9-12 year olds and older adults used a wider pitch range in their conversational speech than 13-14 year olds and young adults. It is possible that the increased pitch range in some participant groups reflects increased engagement with the task resulting in greater animation rather than physiological effects. There was an age effect also in terms of the energy

distribution in the long-term average spectrum of speech. More specifically, we analysed the relative amount of energy in the mid-frequency region of the speech (1-3 kHz) which has been identified as an important predictor of speech intelligibility in background noise. The speech of children had higher mid-frequency energy than young adults (Hazan et al., 2016); this may partly be due to differences in the distribution of spectral energy between child and adult speech. There was also a significant difference between the speech of younger and older adults, with older adults having less energy in the mid-frequency region (Tuomainen & Hazan, 2016b). This could be linked to weaker and more irregular vocal fold excitation in older adults. Speech which has less energy in the mid-frequency region of speech is likely to be more difficult to understand in the presence of noise. This is because this region of the speech spectrum, which contains many acoustic-phonetic cues, is more likely to be masked by noise if it is of lower intensity.

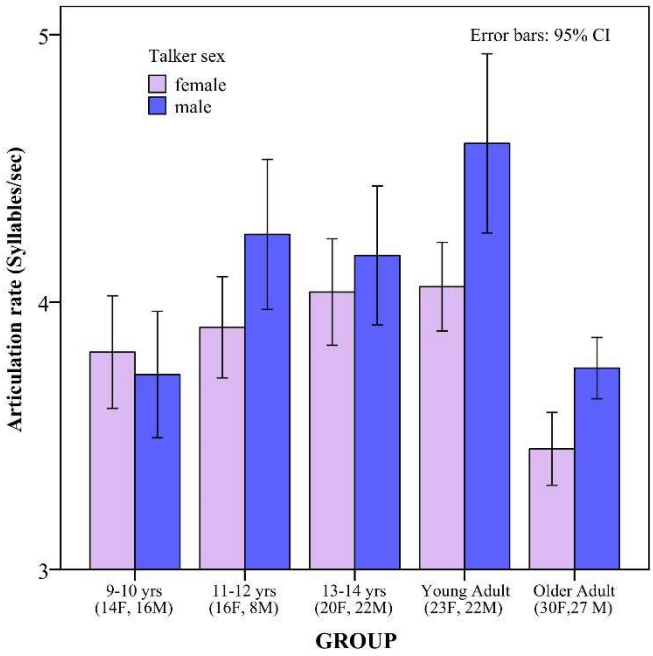


Figure 1.
Conversational articulation rate based on data collected from studies carried out with children (reported in Hazan et al., 2016) and with young and older adults (reported in Tuomainen & Hazan, 2016).

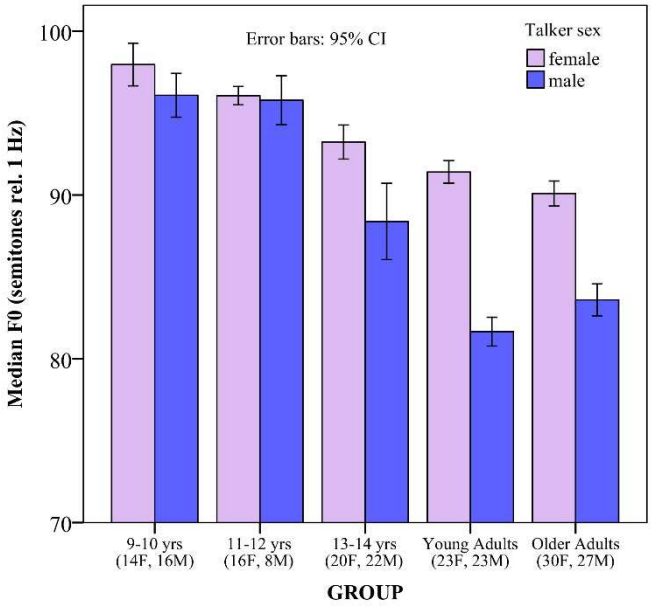


Figure 2.
Median fundamental frequency in conversational speech, based on data collected from studies carried out with children (reported in Hazan et al., 2016) and with young and older adults (reported in Tuomainen & Hazan, 2016).

5 Summary of findings on listener-oriented adaptations in speech production

As the challenging conditions were not identical across the three Diapix studies, and as adaptation strategies can vary as a function of the type of communication barrier to be overcome (e.g., Hazan & Baker, 2011), it is not possible to directly compare the three age groups on a given condition. However, it is possible to compare the adaptation strategies of children to those of young adults for the VOC condition and the strategies of young adults to older adults for the hearing loss simulation condition. Both these conditions proved to cause significant communicative difficulties for participant pairs resulting in longer task transaction times (Hazan et al., 2016). It should be noted also that both involved an interference affecting Talker B only and that the adaptation strategies of Talker A were under scrutiny. These were therefore strategies that were purely listener-oriented and for the benefit of efficient communication and should be differentiated from the type of adaptations (e.g., Lombard speech) that talkers make when directly exposed to interference such as loud background noise.

When comparing the strategies used by children to those used by young adults, it was found that from 9 years of age, children used some adult-like adaptations: they slowed down their articulation rate and increased the mid-frequency energy and median fundamental frequency of their speech for the benefit of their interlocutor. However, unlike adults, 11-14 year olds also increased their fundamental frequency range to counter the effects of vocoding, even though this would not have been transmitted to the conversational partner. This was the case because a noise rather than periodic source was used to excite the vocoder and information about changes in periodicity would therefore not have been present. For child speech only, in the clear speaking style, significant correlations were obtained between increases in mid-frequency energy, reflecting a decrease in spectral tilt, and increases in f0 median and range and, for some child groups, in decreases in articulation rate. Such correlations suggest an increase in vocal effort as would be seen when speaking in a very loud voice or shouting. Children were perhaps using clear speech strategies learnt through their more usual experience of communicating in noisy environments, which would suggest less attunement to the specific characteristics of the interference (Hazan et al., 2016).

When comparing younger and older adults in the hearing loss simulation condition (affecting Talker B), both groups reduced their articulation rate in this condition and both groups also showed increased energy in the mid-frequency region of the long-term average spectrum relative to their spontaneous speech produced in good communicative conditions. Interestingly, for the older adult group with age-related hearing loss only, just as had been found for children, a strong correlation was obtained between increases in fundamental frequency and increases in mid-frequency energy (Hazan & Tuomainen, 2017) as shown in Figure 3. Again, this correlation was totally absent for the young adult group and for older adults with normal hearing thresholds and suggested that some older adults at least were using a strategy of strongly increasing their vocal effort as a clear speech strategy. An intriguing question is why older adults and young children have a greater tendency of strongly raising their voice or shouting when trying to speak clearly for an ‘impaired’ interlocutor. This could be due to a greater degree of frustration experienced when communication problems arise, although there is no objective data to support this hypothesis. It could also be due to a lower degree of inhibition as shouting to an interlocutor may be considered by young adults as inappropriate.

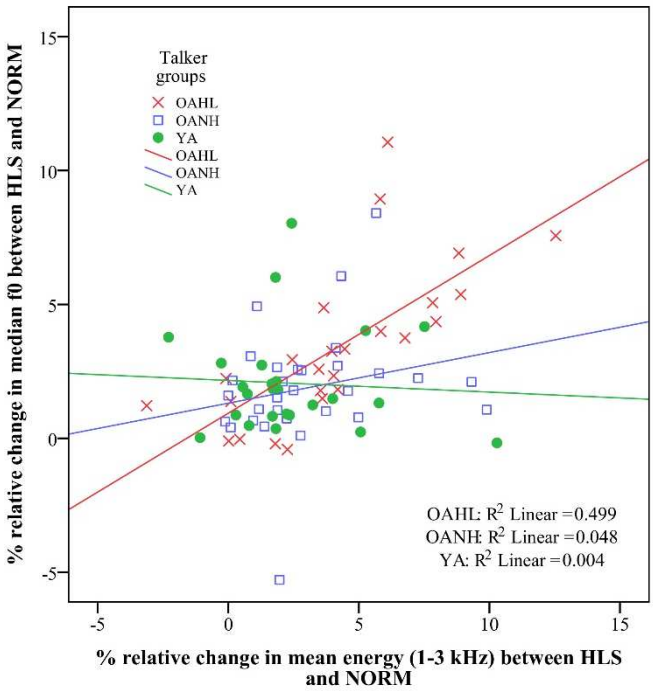


Figure 3.
Scatterplot showing the relation between changes in median f0 and changes in mean energy in the mid-frequency region of the spectrum in the hearing loss simulation (HLS) condition relative to the NORM condition in the study reported in Hazan and Tuomainen (2017) involving young adults (YA), older adults with normal hearing (OANH) and older adults with hearing loss (OAHL).

6 Conclusions

In conclusion, we argue that, despite the increased variability that comes from using spontaneous speech in the analysis of clear speech adaptations, there are benefits in using speech in which these adaptations are naturally elicited due to communicative demands. Further work is needed in order to develop analysis methods that can better represent the dynamic aspects of these adaptations in relation with the degree of communicative success.

References

Anderson, A. H., Bader, M., Bard, E. G., Boyle, E. H., Doherty, G. M., Garrod, S. C., Isard, S. D., Kowtko, J. C., McAllister, J. M., Miller, J., Sotillo, C. F., Thompson, H. S., & Weinert, R. (1991). The HCRC Map Task Corpus. *Language and Speech, 34,* 351-366.

Baker, R., & Hazan, V. (2011). DiapixUK: task materials for the elicitation of multiple spontaneous speech dialogs. *Behavior Research Methods, 43,* 761-770.

Bóna, J. (2014). Temporal characteristics of speech: The effect of age and speech style. *Journal of the Acoustical Society of America Express Letters, 136,* EL116-EL121.

Clark, H., & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition, 22,* 1-39.

Cooke, M., King, S., Garnier, M., & Aubanel, V. (2014). The listening talker: A review of human and algorithmic context-induced modifications of speech. *Computer Speech and Language, 28,* 543-571.

Cooke, M., & Lu, Y. (2010). Spectral and temporal changes to speech produced in the presence of energetic and informational maskers. *Journal of the Acoustical Society of America*, *128,* 2059-2069.

Crawford, M. D., Brown, G. J., Cooke, M. P., & Green, P. D. (1994). The design, collection and annotation of a multi-agent, multi-sensor speech corpus. In *Proceedings of the Institute of Acoustics, 16,* 183-189.

Gósy, M. (2012). BEA − A multifunctional Hungarian spoken language database. *The Phonetician, 105,* 50-61.

Granlund, S., Hazan, V., & Baker, R. (2012). An acoustic-phonetic comparison of the clear speaking styles of late Finnish-English bilinguals. *Journal of Phonetics, 40,* 509-520.

Granlund, S., Hazan, V. L., & Mahon, H. M. (2018). Children's acoustic and linguistic adaptations of peers with hearing impairment. *Journal of Speech, Language, and Hearing Research, 61,* 1055-1069.

Hasan, S.S., Chipara, O., Wu, Y. H., & Aksan, N. (2014). Evaluating auditory contexts and their impacts on hearing aid outcomes with mobile phones. In *Proceedings of the 8th International Conference on Pervasive Computing Technologies for Healthcare* (pp. 126-133).

Hazan, V. L. (2017). Speech communication across the life span. *Acoustics Today, 13,* 36-43.

Hazan, V. L., & Baker, R. (2011). Acoustic-phonetic characteristics of speech produced with communicative intent to counter adverse listening conditions. *Journal of the Acoustical Society of America, 130,* 2139-2152.

Hazan, V., & Tuomainen, O. (2017). Spontaneous speech adaptations in challenging communicative conditions across the lifespan. In *Book of abstracts of Workshop on Challenges in Analysis and Processing of Spontaneous Speech (CAPSS2017)* (3-4).

Hazan, V., Tuomainen, O., & Pettinato, M. (2016). Suprasegmental Characteristics of Spontaneous speech produced in good and challenging communicative conditions by talkers aged 9 to 14 years old. *Journal of Speech, Language, and Hearing Research, 59,* S1596−S1607.

Jacewicz, E., Fox, R. A., & Wei, L. (2010). Between-speaker and within-speaker variation in speech tempo of American English. *The Journal of the Acoustical Society of America, 128,* 839-850.

Kim, M., Horton, W. S., & Bradlow, A. R. (2011). Phonetic convergence in spontaneous conversations as a function of interlocutor language distance. *Laboratory Phonology, 2,* 125-156.

Lam, J., & Tjaden, K. (2013). Intelligibility of Clear Speech: Effect of Instruction. *Journal of Speech Language and Hearing Research, 56,* 2412-2421.

Lecumberri, M. L. G., Cooke, M., & Wester, M. (2017). A bi-directional task-based corpus of learners' conversational speech. *International Journal of Learner Corpus Research, 3,* 175-195.

Lindblom, B. (1990). Explaining phonetic variation: A sketch of the H&H theory. In W. J. Hardcastle, & A. Marchal (Eds.), *Speech production and speech modelling* (pp. 403-439). Dordrecht: Kluwer Academic.

Mattys, S. L., Davis, M. H., Bradlow, A. R., & Scott, S. K. (2012). Speech recognition in adverse conditions: A review. *Language and Cognitive Processes, 27,* 953-978.

McInerney, M., & Walden, P. (2013). Evaluating the use of an assistive listening device for communication efficiency using the Diapix task: A pilot study. *Folia Phoniatrica Logopedia, 65,* 25-31.

Murfitt, T., & McAllister, J. (2001). Comprehension of spoken descriptions by novel listeners in monologue and dialogue. *Language and Speech, 44,* 325-350.

Nip, I. S. B., & Green, J. R. (2013). Cognitive and linguistic processing primarily account for increases in speaking rate with age. *Child Development, 84,* 1324-1337.

Pettinato, M., Tuomainen, O., Granlund, S., & Hazan, V. L. (2016). Vowel space area in later childhood and adolescence: effects of age, sex and ease of communication. *Journal of Phonetics, 54,* 1-14.

Scarborough, R., Dmitrieva, O., Hall-Lew, L., Zhao, Y., & Brenier, J. (2007). An acoustic study of real and imagined foreigner-directed speech. *Journal of the Acoustical Society of America, 121,* 3044-3044.

Scarborough, R., & Zellou, G. (2013). Continua of clarity: "clear" speech authenticity and lexical neighborhood density effects in production and perception. *Journal of the Acoustical Society of America, 134,* 3793-3807.

Smiljanic, R., & Bradlow, A. (2009). Speaking and hearing clearly: talker and listener factors in speaking style changes. *Language and Linguistics Compass, 3,* 236-264.

Solanki, V., Stuart-Smith, J., Smith, R., & Vinciarelli, A. (2015). Measuring mimicry in task-oriented conversations: the more the task is difficult, the more we mimick our interlocutors. In *Proceedings of InterSpeech 2015* (pp. 1815-1819).

Sørensen, J., Fereczkowski, M., & MacDonald, E. N. (2017). The effect of noise and second language on turn taking in task-oriented dialog. *Journal of the Acoustical Society of America, 141,* 3520.

Stamp, R., Schembri, A., Evans, B. G., & Cormier, K. (2016). Regional Sign Language Varieties in Contact: Investigating Patterns of Accommodation. *Journal of Deaf Studies and Deaf Education, 21,* 70-82.

Tuomainen, O., & Hazan, V. L. (2016a). Articulation rate in adverse listening conditions in younger and older adults. In *Proceedings of Interspeech 2016* (pp. 2105-2109).

Tuomainen, O., & Hazan, V. (2016b). Suprasegmental characteristics of spontaneous speech produced in good and challenging communicative conditions by younger and older adults. *Journal of the Acoustical Society of America, 140,* 3444.

Van Engen, K. J., Baese-Berk, M., Baker, R. E., Choi, A., Kim, M., & Bradlow, A. R. (2010). The Wildcat Corpus of Native- and Foreign-Accented English: communicative efficiency across conversational dyads with varying language alignment profiles. *Language and Speech, 53,* 510-540.

White, L., Wiget, L., Rauch, O., & Mattys, S. L. (2010). Segmentation cues in spontaneous and read speech. In *Proceedings of the 5th Conference on Speech Prosody* 2010 (pp. 1-4). Chicago.

Zurek, P. M., & Desloge, J. G. (2007). Hearing loss and prosthesis simulation in audiology. *Hearing Journal, 60,* 32-33, 36, 38.

ASPECTS OF COARTICULATION

Vesna MILDNER
Department of Phonetics, Faculty of Humanities and Social Sciences,
University of Zagreb, Croatia
vmildner@ffzg.hr

Abstract
The paper provides a summary of various types and aspects of coarticulation. After setting a framework that includes general considerations such as biomechanical and language-specific issues, the distinction between anticipatory and carry-over coarticulation, the discussion of articulatory pressure/resistance and its scope, it analyzes different levels at which coarticulation occurs: lips, tongue, velum and larynx. The review of the most influential models and theories from the 1960s until the present reveals that a comprehensive explanation of coarticulation is yet to be offered. In terms of neuromotor control, it shows that very little research has been done specifically on coarticulation, so most conclusions in available literature are indirectly derived from studies of speech production in general. The paper also tries to shed some light on coarticulation in populations that have been studied less extensively, such as children and clinical cases. The goal of this review is to give a brief overview of the current 'state of affairs' in coarticulation studies and argue for the need to extend them to more languages, less than typical populations and to higher levels of processing.

Keywords: coarticulation, coarticulatory resistance, coarticulatory pressure, speech acquisition, impaired speech motor control.

1 Introduction

Speech consists of segmental and suprasegmental features. However, the term 'segmental' may be a little misleading. It refers to individual sounds (rather than superimposed or accompanying characteristics of the utterance, i.e., prosody = suprasegmentals) and it may be inferred that the segments follow each other in an orderly fashion – one appearing after another has been completed. This misconception is readily recognized when one attempts to analyze a sample of speech and, as a first step, tries to separate/distinguish the individual sounds. It becomes painfully obvious that it is very difficult (if not impossible) to decide with certainty where one segment ends and the following one begins. The culprit is coarticulation.

In most general terms, coarticulation is defined as the influence that speech sounds exert upon one another in running speech.

DOI: http://doi.org/10.18135/CAPSS.27

Although the terms coarticulation and assimilation are frequently used interchangeably and they both occur as consequences of sound context, the distinction between them is commonly described as assimilation referring to audible change resulting in the perception of another phoneme and coarticulation being reserved for the physiological domain of speech organs coordination (Hardcastle & Tjaden, 2011). In terms of generative phonology, it can be said that assimilation belongs to the realm of linguistic competence and coarticulation to that of performance (Chomsky & Halle, 1968). The former is language-determined (i. e., governed by language-specific rules, e. g., phonological ones) while the latter is universal although it may appear to differ across languages, particularly in degree, and what is considered assimilation in one language may be described as coarticulation in another (e.g., vowel harmony vs., transconsonantal coarticulation) (Farnetani & Recasens, 2010; Volenec, 2015; Horga & Liker, 2016). There have been other attempts to define this distinction, including the suggestion of a listener-centered approach (Fowler, 1980), but the issue is far from being resolved and it involves the age-old discussion about the phonetics vs. phonology domains.

In the literature, there are a number of terms referring to what is called coarticulation in this text: coordination, gestural overlap, interarticulator timing, context effects, sound-transitional effects (to name just a few) (Hardcastle & Tjaden, 2011).

2 General aspects of coarticulation

Coarticulation has its biomechanical and language-specific aspects. The biomechanical aspect is supposedly universal, because it is in essence a manifestation or consequence of characteristics and functioning of our speech production system. It is not limited to inertia of articulators, but rather it involves continuous modifications/adjustments in line with communicative demands during speech (not necessarily dependent on speech tempo). On the other hand, language-specific aspects of coarticulation are apparent in the assumption that coarticulation is governed by language rules, and therefore not a mere consequence of what our speech organs can or cannot do. If coarticulation were JUST biomechanically determined, the levels of speech planning/programming and speech execution would be independent without the possibility of feedback and error management. It seems appropriate to view coarticulation as a combination of the two aspects at an undetermined ratio, and here is where the distinction between anticipatory and carry-over coarticulation solves at least part of the problem.

In **anticipatory** coarticulation (also called forward, regressive, right-to-left) the current sound is influenced by the one following it, i.e., its place of articulation is slightly modified and approximates the place of articulation of the succeeding sound (e.g., the /g/ in goon is produced with the tongue dorsum in a

more backward position than in geek). It is considered to be a sign of speech (motor) planning, and language-determined. In other words, it is a higher level process whose patterns vary across languages (Keating & Lahiri, 1993). It is therefore susceptible to disruption in speech disorders characterized by impaired (speech) motor control. From the perceptual viewpoint it contributes to faster and more accurate perception due to the fact that acoustic cues of the incoming segment are present in the current one (e. g., Dahan et al., 2001; Salverda et al., 2014). Experiments combining eye-tracking with cross-splicing of initial CV + final C from minimal pairs differing in the final C (e.g., net vs. neck) revealed that about 200 ms after word onset, i. e., before the actual articulation of the final consonant, subjects relied on the information contained in the vowel which was actually derived from the spliced out consonant: when [t] from net was added onto [ne] from neck, the subjects favored the picture of a neck before the [t] was reached (Dahan et al., 2001). This impact of coarticulation in lexical decision tasks has been found in experiments with nasalization as well (Beddor et al., 2013).

In **carry-over** coarticulation (also referred to as perseverative, backward, left-to-right, retentive) the current sound influences production of the following one (e. g., in boots lip protrusion necessary for the production of /u/ carries over onto /t/ and /s/), and it is generally taken as a consequence of inertia of the speech production apparatus, i.e., it is biophysiologically determined, and thus universal, although it has been argued that it involves a certain degree of planning as well (Recasens, 1999).

One and the same sound in a sequence may have both anticipatory and carry-over effects. For example, in the word /ana/ the /n/ has anticipatory effect on the initial /a/ and carry-over effect on the final /a/, as suggested by acoustic record, oral and nasal flow curves, and synchronized EPG, corresponding to opening of the velopharyngeal port right after onset of the initial /a/ and its remaining open until the end of the final /a/ (Farnetani & Recasens, 2010).

The range/scope and direction of coarticulatory effects is determined by a set of constraints that may include physiological features of the articulators (→ resistance), suprasegmental features (stress patterns, prosodic and syntactic boundaries, syntactic structure, rate of articulation, clarity, speech style) and language specific constraints – phonological structure (Hardcastle & Tjaden, 2011).

Coarticulatory resistance and coarticulatory pressure (dominance, aggressiveness) are two properties of sounds that are positively correlated: phonetic segments that are especially resistant to coarticulatory effects from the adjacent segments exert maximal coarticulation on them.

There is general agreement among researchers that there are some parts of the speech signal that are more resistant to coarticulatory effects and exhibit a higher degree of invariance. Krull (1989) reported that labial consonants are more affected than dental ones by coarticulation, and that in CVC syllables the vowel

exhibits greater anticipatory coarticulation on the preceding consonant than carry-over effect on the following one. Interestingly, in contrastive hyperarticulation voiced and voiceless stops are affected differently: in order to avoid ambiguity, speakers decrease VOT in voiced stops and increase it in voiceless ones (Mücke et al., 2017). Liker and Gibbon (2018) report the tendency of /z/ to be more resistant to coarticulation effects than /s/. The adaptation of tongue position in lingual consonants to the tongue position of an adjacent sound is constrained by intra-articulator coordination and coupling of tongue dorsum with its other parts (i.e., lamina and tip): for example, in the production of postalveolars, such as /ʃ/, the tongue dorsum is critical and this limits their potential for adaptation to the subsequent vowel because tongue dorsum is slow and inert, therefore more resistant to adaptation than e.g., /s/ where the tip of the tongue is active. Alveolars closely follow in the degree of constraint, and labials (e.g., /p/) are more affected than alveolars (Zharkova et al., 2015). Coarticulatory resistance and pressure are further discussed below, in the context of DAC model.

Iskarous et al. (2013) propose a coarticulation-invariance scale on which the amount of 'mutual information', i.e., information shared by candidates for coarticulation, is proportional to the degree of coarticulation. The amount of information is based on the measurements of physical positions of articulators during speech production and it is high in coarticulation and low in invariant condition. Testing their scale on American English, Catalan and German data revealed that it confirms the previous empirical studies of contextual (in)dependence of specific sound categories, i.e. articulatory resistance.

There have been claims that coarticulation spreads over up to 6 neighboring sounds, but the span of coarticulation is still an issue for debate (Kent & Minifie, 1977; Farnetani & Recasens, 2010). Also it has been suggested that it varies across coarticulatory systems, labial coarticulation having the largest span, followed by velar and lingual coarticulation (for review see Volenec, 2015). Bell-Berti and Harris (1975) suggested that carry-over effects are more extensive than the anticipatory ones. However, regardless of the actual numbers, the finding that several units are in various stages of planning, adjustment, execution and somatosensory feedback at the same time, has implications for understanding the system of motor control.

Coarticulation is not limited to word level – in connected speech it is present across word boundaries as well. For example, in producing the noun phrases lean bacon or green boat, the alveolar nasal place of articulation of /n/ moves toward the bilabial place of articulation in anticipation of the bilabial /b/. Salverda et al. (2014) have shown that listeners make immediate use of anticipatory coarticulation in the determiner to predict the initial sound(s) of the upcoming word (in a paradigm where the determiner the is followed by targets starting with different consonants), which can be explained by the finding that

from its very onset, the neutral schwa [ə] exhibited strong influence of the following sound, as shown by F_1, F_2 and F_3 trajectories.

Recasens (2015) reports the effect of stress and speech-rate variations on overall vowel duration, second formant frequency and coarticulation size but not on the consonant-specific patterns of degree and direction of vowel coarticulation, and interprets these results as indication that coarticulatory changes caused by prosody conform to the basic principles of segmental coarticulatory organization. Cho et al. (2017) found that prominence enhanced nasality of the consonant and orality of the vowel (rather than nasality) showing the coarticulatory resistance to nasal effects, even when the focus was on the nasal. Boundary strength induced prosodic position-dependent contradictory patterns. They conclude that vowel nasalization is under speaker's control and take their results as evidence of close relationship between the dynamics of speech timing and (the need to preserve) linguistic contrasts.

3 Levels/systems of coarticulation

Most research focuses on tongue related coarticulation, but coarticulatory processes are present at the laryngeal, nasal and labial levels as well. In other words, articulators studied in coarticulation are lips, tongue, velum and larynx. Mandibular movements are typically observed together with lips and tongue because they are considered to be integral part of changes in the position of the two.

Processes associated with **lips** are usually referred to as lip rounding, spreading or protrusion (e.g., in the word choose under the influence of /u/ lip rounding will begin during production of /tʃ/, and in cheese lips will be spread during production of /tʃ/ under the influence of /i/) and their acoustic aspects are described in terms of formant changes. Full description of lip aperture requires both the horizontal and vertical axis specification: rounded sounds have smaller aperture along both axes. In languages in which both rounded and unrounded vowels are constituents of the phonemic repertoire (e.g., Swedish, German, French), roundedness is associated with more complex articulatory characteristics, it is less variable and more resistant to coarticulation than in languages in which it plays no phonologically distinctive role (Farnetani & Recasens, 2010; Horga & Liker, 2016). Labial coarticulation seems to have the largest scope - Swedish electromyographic data reveal lip rounding starting up to 600 ms before the actual rounded vowel (Lubker et al., 1975, as cited in Volenec, 2015).

Coarticulatory displacements of the **tongue** along the horizontal (front – back) and/or the vertical axis (high/close – low/open) also result in corresponding formant shifts (e.g., in /aga/ and /igi/, the tongue shape during the closure for the /g/ is a blend of the gestures for the vowels and the consonant). It is important to note that tongue tip and tongue body may be controlled independently. This is also one of developmental constraints of articulation, since it seems that children

take longer to master this selective control and to replace movements of the tongue as a whole with independent control of its tip/blade and body. Zharkova et al. (2012) contribute absence of significant effects on /s/ by the following vowel (particularly /i/ and /u/) in children, compared with adults, to this lack of differential control. They also relate their interpretation to Cheng et al. (2007)'s study and conclude that such differential coordination occurs around 9 years of age and is further refined into late adolescence; according to Schötz et al. (2013) it may extend even into the late 20s.

Lowering of the **velum** (typically referred to as nasalization) has repercussions in changes of the oral formant structure and occurrence of nasal formants (e.g., in the word dance, the nasal /n/ may initiate velum lowering as early as the initial /d/, causing nasalization of the oral /æ/). Coarticulatory activity between nasal consonants and neighboring vowels is a two-way street, i.e., reciprocal: during vowel production the velum is in a lower position in the vicinity of nasal consonants than in the vicinity of the non-nasal ones, and during nasal consonant production, the vicinity of close vowels results in lower velum position than the vicinity of open ones (Farnetani & Recasens, 2010; Horga & Liker, 2016). Nasal coarticulation is both language-determined and physiological, its extent depends on the phonemic repertoire of the language, and Bouchard and Chang (2014) suggest it is under speaker's control.

Coarticulation at the level of **larynx** is associated with vocal fold activity and the presence or absence of periodicity, but it is also directly related to the levels above the larynx. The degree and duration of laryngeal coarticulation are affected by place and manner of articulation. For example, the opening of the glottis in the articulatory process of devoicing has been reported to start earlier in fricatives than in stops (Hoole, 1999) and VOT has been found to vary across places of occlusion (Bakran, 1993; Horga & Liker, 2016). Similarly, different consonantal contexts (i.e., fricatives vs. stops) affect laryngeal activity during vowel production in different ways, both with respect to variability and timing. Research into correlation between laryngeal and lingual places of articulation has yielded inconclusive and often contradictory results largely due to different research questions and methods (for a review, see Horga & Liker, 2016; Liker & Gibbon, 2018).

4 Models

Several models have been developed to account for coarticulation. They include the look-ahead, articulatory syllable, time-locked, window, coproduction and articulatory phonology models (for a more extensive discussion of these and other models, see Farnetani & Recasens, 2010; Volenec, 2015; Horga & Liker, 2016).

The target undershoot model (Lindblom, 1963), although not explicitly a model of coarticulation, posits that articulators frequently fall short of their

target (hence, undershoot) due to responding to simultaneous articulatory commands, but the relationship between the target and its mental representation is not clearly defined. According to Lindblom, the degree of coarticulation is a manifestation of speech economy; however, it does not depend exclusively on speech rate (as posited in earlier works), but on the demands for perceptual contrast and style (Lindblom, 1990; Moon & Lindblom, 1994). The strategies speakers use are determined by duration, input articulatory force and time constant of the system. Lindblom (1963) also proposed an elegant tool for measuring coarticulation, locus equation, which has been shown across many languages to be a robust indicator of its degree (Sussman et al., 1993; Bakran & Mildner, 1995). Locus equations are linear regressions of the onset of F_2 transition on F_2 target (at the vowel nucleus). The calculated slope and intercept depend on the consonant place of articulation. In CV syllables, steeper slopes are indicative of higher degree of coarticulation. Based on her comparison of locus equation and EPG data for English CV syllables, Tabain (2000) also suggested that a distinction should be made among consonant categories. Namely, her data revealed that alveolar and velar stops and nasals exhibit a good correlation between locus equation and coarticulation, as opposed to fricatives (especially /z/ and /ð/), where the correlation between EPG and locus equation data was very poor (possibly due to fricative noise obscuring the F_2 transition, and/or locus equation being incapable of encoding the more subtle differences in the degree of coarticulation found in coronals. However, according to Löfqvist (1999), EMA data do not support the notion that the slope in locus equation approach is indicative of the degree of CV coarticulation.

In the articulatory syllable model (Kozhevnikov & Chistovich, 1965, as cited in Farnetani & Recasens, 2010) coarticulation is limited to within CV sequences, which in light of prevailing evidence from various languages is too limited a scope.

Feature-spreading (sharing) / look-ahead model: Henke (1966) proposed a computer model positing that a segment (i.e., input from the neural representation level) will have coarticulatory effects that start as early as possible if there are no contradictory specifications. Along these lines Daniloff and Hammarberg (1973) proposed that phonetic representation includes articulatory and coarticulatory specifications, and the model scans upcoming units (i.e., looks ahead) for specified feature values, all with the goal of achieving smooth transitions between segments. However, empirical data across various languages have shown that contradictorily specified adjoining segments may still be subject to coarticulation, and that unspecified segments may have some resistance to coarticulation and/or may behave differently in different contexts (Farnetani & Recasens, 2010). This indicates that phonological features and their specifications are too rough units and coarticulation needs to be

defined in much finer terms that should include articulatory, aerodynamic, acoustic and perceptual constraints.

To account for the gradual changes in the process of coarticulation, Keating (1990) proposed the so-called window model of coarticulation, according to which spatial and temporal context-dependent variations are governed by phonetic rules of the grammar. A window is the point at which categorical phonological input is converted into non-binary phonetic description and its size for each feature is positively correlated with variation, and in turn, negatively with specificity, e.g., poorly specified features are associated with wide windows and are subject to a high degree of contextual variation, hence coarticulation. On the other hand, narrow windows correspond to greater coarticulatory resistance. Windows are connected by paths (interpolation curves) that reflect articulatory and/or acoustic variations over time. Window size and position taken together with the shape of the path (contour) determine coarticulation, governed by demands for smoothness and least articulatory effort. Languages differ in coarticulation due to differences in phonology or phonetics. Major arguments against the model are based on experimental findings that failed to corroborate (expected) direct interpolation in contexts of unspecified sounds and on claims that the model is too simplified and does not account for the complex nature of speech production (Farnetani & Recasens, 2010; Volenec, 2015; Horga & Liker, 2016, and references therein).

The concept of articulatory gestures (not to be confused with articulatory movement or articulatory target) is associated with the task-dynamic model of speech production, in which phonetic gestures rather than phonological features or segments are inputs to the process of production (and by extension, coarticulation) (Fowler, 1980; Fowler & Saltzman, 1993; for discussion, see Farnetani & Recasens, 2010; Volenec, 2015; Horga & Liker, 2016). Articulatory gestures are defined as target/goal-oriented, serially ordered planned actions (of all articulators involved in the production of a particular sound), with intrinsic temporal structure, and context-independent. The speed at which this internal (re)organization takes place in cases of changed circumstances (e.g., obstruction, damage to articulators) indicates that it is not centrally controlled (Löfqvist (2010) proposes brainstem as the crucial point of integration of incoming somatosensory feedback and motor control, see below). In this context, coarticulation is seen as coproduction of articulatory gestures, i.e. overlap of neighboring ones. The extent of overlap depends on speech tempo and articulatory conditions and is, as a rule, controlled at the level of planning. When two articulatory gestures ‘compete’ for involvement of the same articulators, the result will depend on the strength of the two gestures: when their strengths are similar their influence will average out, otherwise the stronger one will suppress the effect of the weaker one. In other words, the stronger one can be character-

rized as having greater coarticulatory resistance and, accordingly, greater coarticulatory effect. Cross-linguistic differences can be attributed to different gestural organization.

One of the models relying on articulatory gestures is the so called time-locked model of anticipatory coarticulation, which posits that component gestures of a segment begin a fixed interval of time before the phonetic target is achieved. However, not all experimental data support this model and some lend preference to look-ahead models. Also, cross-linguistic comparisons reveal a great deal of variability among languages (for discussion, see Farnetani & Recasens, 2010).

Recasens et al. (1997) proposed the Degree of articulatory constraint (DAC) model, according to which coarticulation is a process that continuously involves more than one speech unit. The model is based on Catalan data and focuses on lingual coarticulation (which is its major limitation). It postulates that the three elements of coarticulation: degree, temporal extent and direction are determined by the requirements imposed on the tongue in the process of speech production. Vowels and consonants are assigned values – the higher the value (degree of articulatory constraint) the more resistant the sound is to coarticulatory effects and the greater its coarticulatory pressure with respect to adjacent segments. Consonants requiring a high degree of articulatory precision (e.g., alveolar trill) have the highest DAC value, and those that do not require a great amount of tongue body activity (e.g., labials) have the lowest. Similarly, within vowels, the ones requiring the greatest amount of tongue dorsum displacement (e.g., front vowels) have the highest DAC value as opposed to those with an unspecified target (e.g., [ə]). The temporal extent is determined by the articulatory constraint in such a way that anticipatory effects start earlier when the preceding sound is relatively unconstrained (suggesting that it is not exclusively a result of planning). Carry-over effects are more variable and may take longer. Within the model, vowels and consonants tend to favor anticipatory or carry-over direction (e.g., dark /l/ favors anticipation, /ɲ/ favors carry-over component) (more on this in Farnetani & Recasens, 2010). Additionally, highly constrained consonants do not exhibit coarticulatory effects determined by their position within the syllable.

Current theories and models fail to offer a comprehensive explanation of coarticulation (i.e., at all the levels it occurs), and to account for cross-linguistic differences. Moreover, they seem to base their assumptions on differently defined domain/origin, function and control of articulation (Farnetani & Recasens, 2010; Horga & Liker, 2016).

5 Neuromotor control

At the level of neuromotor planning and programs Gracco and Löfqvist (1994) suggest that speech movements are organized into aggregates consisting of several functionally related articulators. These aggregates correspond to

articulatory gestures. Each sound has its neuromotor representation based on the muscles that need to be activated for its production and their spatial and temporal coordination. These structures correspond to neurobiological equivalents of the phoneme. Obviously, as summarized by Kent and Minifie (1977), in the process of turning phonemic representations into actual speech the mentally stored discrete and invariant units undergo modifications not only in their boundaries but in their acoustic and articulatory properties as well. Coarticulation requires / relies on additional adaptive processes (central neural mechanisms) associated with these representations which enable combinations into larger sequences, so that the underlying units are modified in actual production. This requires that neural control of the units be flexible enough to allow for contextual variations. At the speech perception end, this flexibility is manifested in the ability to process the message successfully in spite of the lack of invariance present in the speech signal, and the circle back to matching the 'ideal' representations is complete.

Cerebral areas involved in speech processes have been studied extensively for decades, but relatively recently have some areas other than the cortex been recognized as important in speech motor planning and execution, such as the left insula (Dronkers, 1996), cerebellum (Gordon, 1996; De Smet et al., 2007) and thalamus (for review, see Katz, 2000). Also, with the discovery in the 1990s and subsequent fruitful research of mirror neurons (Rizzolatti & Craighero, 2004) the cooperation of sensory and motor networks has received the attention it deserves because of its implications for understanding speech production and perception processes (among others) and language in general.

Based on fMRI data of eight healthy volunteers, Riecker et al. (2005) suggest two levels of speech motor control associated with motor preparation (medial and dorsolateral premotor cortex, anterior insula and superior cerebellum) and execution processes (sensorimotor cortex, basal ganglia and inferior cerebellum). Brendel et al. (2010) on the basis of clinical data further elaborate on this speech motor control network and suggest organization into (at least) three functional neuroanatomical subsystems: one devoted to planning of movement sequences (premotor ventrolateral-frontal cortex and/or anterior insula), one being activated in the process of preparing for or initiation of upcoming verbal utterances (supplementary motor area), and the third being in charge of execution (corticobulbar system, basal ganglia, cerebellum). An interesting finding of their fMRI study was that the timing of activation of these different neural circuits was not fixed, but changed as the task progressed and one might assume that this flexibility is advantageous to accommodating and adjusting accordingly to feedback information coming from somatosensory networks. Such fluctuations in alternately activating various language-associated

areas (predominantly in the left hemisphere) were reported by Nakai et al. (2017) as well.

Löfqvist (2010) describes neural motor control of speech as a distributed network consisting of neuronal circuits and centers at different levels. Within the network there is communication between the periphery and a central/executive unit that receives and processes incoming information about the current situation/context, and based on that selects not only the appropriate muscles to be activated but also determines/adjusts the level of their involvement and spatio-temporal organization/coordination. As in gross motor activity, the brainstem seems to be crucial for such integration involved in articulation as well. The rapid, functional compensations following perturbations to articulators are in agreement with such a distributed system.

Sensorimotor integration is an important part of this process and the perceptual system has a crucial role in the self-monitoring of speech (Hickok et al., 2011; Bouchard & Chang, 2014). Consequently, speech production requires activation of internal representations of sensory speech targets in addition to motor speech representations. Error signals that are received in the process of speech may be dealt with in two ways: by modifying motor programs for immediate target attainment, or by modifying representations for future reference (Hickok et al., 2011). Bouchard & Chang (2014) found significant activity in the ventral sensorimotor cortex (vSMC) during production of CV syllables. This activity robustly predicted acoustic parameters across vowel categories and different renditions of the same vowel. They also found significant contextual effects on vSMC representations of produced phonemes, which they took as indication of active control of coarticulation. In terms of direction, they found that representations of vowels were biased toward the representations of the preceding consonant, and representations of consonants were biased toward subsequent vowels.

Broca's area (inferior frontal gyrus in the language-dominant, usually left, hemisphere) has of course been described as undisputed crucial area in speech production (and a number of other language-related functions). Consequences of damage to this area include impaired motor planning of speech articulation seen as unsuccessful attempts at reaching the target while producing polysyllabic words and maintenance of serial order of phonemes (Davis et al., 2008). Peeva et al. (2010) used fMRI to study representation of speech segments of varying complexity and found that the left medial premotor regions process phonemes (while syllables are processed in the left lateral premotor regions) and that these areas have projections to primary motor cortex along which representations are transformed into motor commands to the articulators, thus confirming the dominance of the left (pre)motor cortex in speech planning and initiation.

By recording the activity of the lateral superior temporal cortex, Leonard et al. (2015) examined auditory processing of sound sequences (words and nonwords) and reported data that support the interactive bottom-up and top-down processes, i.e. integration of physical stimulus characteristics (bottom-up) with their contextual sequential structure and subconscious phoneme sequence statistics and higher-order linguistic knowledge (top-down). Moreover, their subjects' neural responses revealed dynamical encoding of language-level probability of preceding and upcoming sounds, clearly showing correspondence with phoneme onsets and transitions. This may help explain how even high degrees of coarticulation do not cause perceptual break-down and confirms the importance of superior temporal cortex in processing language stimuli and sensorimotor integration.

6 Developmental aspects

Speech planning and production are governed by developmental processes. While it is clear that this is reflected in coarticulation, exactly how it is manifested is less unambiguous. Some authors suggest that children's linguistic units are larger and less specified than adults' (hence, characterized by more variation and coarticulation) and that they become less coarticulated as refinement of speech production proceeds throughout maturation and mastering the language (e.g., Kent et al., 1996; Nittrouer et al., 1996). Nittrouer and Whalen (1989) report greater evidence of coarticulation in the fricative-vowel syllables of children than in those of adults. This increased coarticulation led to improved vowel recognition from the fricative noise alone, indicating that the coarticulated sound can be identified without correct identification of the most prominently specified one.

On the other hand, some authors (e.g., Cheng et al., 2007; Zharkova et al., 2012) report vowel contexts of greater coarticulatory influence on speech production in adults than in children (with children having greater within-speaker variability in the degree of coarticulation) and attribute that to children's immature speech, characterized by insufficient coordination and motor control, which is particularly apparent in children younger than nine years.

Sereno et al. (1987) claim that both the acoustic and the perceptual data show strong anticipatory labial coarticulation for the adults and comparable, although less consistent, coarticulation in the speech stimuli of the children. Based on acoustic and video data, Katz et al. (1991) conclude that young children and adults produce similar labial and lingual (sV) anticipatory coarticulatory patterns, but also (based on perceptual data) that coarticulatory cues in the speech of their 3-year-olds are less perceptible than those of older children or adults. They attribute the latter finding to the possibility that, fricatives being (among) the most difficult sound categories to master, children as young as 3 years, have less precise articulation of

/s/. They also found that children show greater variability than adults, but not greater degree of intrasyllabic coarticulation.

At least some of the differences among studies may be attributable to different environments studied and methods applied. A clearer picture of developmental aspects of coarticulation emerges if a distinction is made among coarticulation contexts: e.g., labial coarticulation seems to mature earlier than lingual (Katz & Bharadwaj, 2001; Goffman et al., 2008). However, there appears to be general agreement that (co)articulation is more variable in children than in adults and that stability increases with age (Cheng et al., 2007; Zharkova et al., 2011, 2012, 2017). This can be explained by (at least) two factors: general cognitive maturation (which takes care of speech planning), and practice due to experience (which takes care of sensory-motor precision). However, Schötz et al. (2013) claim that age related changes in speech motor control may not be complete before the age 30.

7 Clinical aspects

Since coarticulation is a marker of fine motor control in speech production it is an important issue in studies of motor disorders that have consequences on speech output, primarily its intelligibility, e.g., apraxia of speech, dysarthria in Parkinson's disease (PD).

Less coarticulation in clinical populations may be a direct consequence of the disorder, but it may also be an indirect result of slower speech rates that are frequently found in such subjects. Additionally, many of these patients produce speech movements that are reduced in size or amplitude (Hardcastle & Tjaden, 2011). One such example is smaller area of the vowel space (as determined by F_1 & F_2 values) but also a trade-off between correctness and time necessary to produce affricates found in hearing impaired speakers (e.g., Liker et al., 2007; Mildner & Liker, 2008).

Tjaden (2000) compared speech rate effects on coarticulation in PD patients and healthy subjects and found that coarticulation tended to increase with faster rates and decrease with slower rates, but more systematically so in control speakers. Overall results suggest increased coarticulation in PD patients relative to control speakers. This effect was not entirely attributable to the more rapid speaking rates for speakers with PD.

Dysarthria is characterized by impaired speech production manifested as impaired rate, intonation, articulation, volume, voice quality and nasality, as a consequence of damage to basal ganglia, thalamus, cerebellum or cerebral cortex (Chang et al., 2009). It typically accompanies various neurological impairments such as ALS, PD or traumatic brain injury, but it does not necessarily cause perceptually meaningful deficits in articulation and coordination. Some studies report normal coarticulatory patterns, some reveal subtle changes (increase or

decrease of context effects), and yet others suggest different patterns at different levels (e.g., normal supraglottal coordination but incoordination at the laryngeal-supralaryngeal interface/level causing difficulties in stopping vocal fold vibration at the transition from a voiced to a voiceless sound) (for discussion, see Hardcastle & Tjaden, 2011).

Deep brain stimulation (DBS) has been shown to improve articulation in dysarthric PD patients by inducing changes in fine motor control. Sauvageau et al. (2014) have examined the influence of bilateral subthalamic nucleus DBS on carry-over coarticulation in CV combinations. Even though the consonant context influenced vowel articulation, this coarticulatory phenomenon did not vary as a function of the DBS across their 8 PD patients. In a previous study Wang et al. (2006) found that the side of DBS had different effects on speech production – left-hemisphere stimulation altered articulation accuracy. With right-hemisphere stimulation it remained unchanged or improved. However, not all studies report speech improvement and there is great variability among studies and patients (Aldridge et al., 2016).

Apraxia of speech (AOS) is manifested as impaired speech motor planning (especially for complex syllables) and has been associated with damage to the left anterior insula (Dronkers, 1996) and with damage to the posterior inferior frontal gyrus (Hills et al., 2004). However, it must be stressed that a strictly localizationist approach is not justified. AOS is often, but not always associated with Broca's aphasia (Katz, 2000).

In AOS, VOT values for voiced and voiceless stops tend to overlap (even when consonants are perceived as correct), and there is great variability in VOT for the same stop. These two features are taken as evidence of poor coordination of laryngeal-supralaryngeal events, and in turn this is interpreted as AOS affecting timing or coordination between articulators. This is corroborated by studies of anticipatory coarticulation (Katz, 2000; Hardcastle & Tjaden, 2011) revealing great variability in timing (especially in labial and lingual coarticulation). In addition to increased variability, some studies report delays in coarticulation: Patients with AOS begin vowel gesture in CV syllables later than controls. Ziegler and von Cramon (1986) attribute lack of coarticulatory cohesion in the speech of a patient suffering from verbal apraxia to a consistent delay in the initiation of anticipatory vowel gestures.

Studies of coarticulation in AOS and cerebellar ataxia suggest that anticipatory coarticulation has a multifocal representation in the nervous system and perserveratory coarticulation is regulated, at least in part, by the cerebellum (Katz, 2000).

In fluent types of aphasia, e.g., Wernicke's, anticipatory coarticulation is preserved, as evidenced by perceptual-acoustic studies. However, not all inconsistencies in (co)articulation and coordination that are physically present

are perceptually noticeable as revealed by EPG data of an anomic aphasic (Hardcastle & Tjaden, 2011). For example, abnormal prevoicing and nasalization have been reported in Wernicke's patients (Katz, 2000).

Common speech characteristics of childhood/developmental apraxia of speech (DAS) are numerous and inconsistent consonant errors and context-related substitutions, groping and overall poor intelligibility. On the basis of locus equation calculations Sussman et al. (2000) concluded that reduced intelligibility of children with childhood apraxia of speech may be attributed to their inability to sufficiently distinguish among stop place categories due to poor refinement of coarticulation levels. Studies of DAS have also revealed inconclusive results with respect to coarticulation – some report earlier and stronger anticipatory vowel effects, some just the opposite. Usually children with DAS exhibit more inter- and intra-subject variable patterns than children with typical speech acquisition, and their speech suggests deficits in motor planning as well as in syllabic programming. Apparently, the breakdown occurs during the transformation of phonological representation into articulatory (motor) program; however it seems to involve not just execution but also the acquisition and automatization of a speech production plan (Maasen et al., 2001). This is also supported by the study of Grigos and Case (2017) where the effect of practice was found in both typically developing children and children with DAS, but while the former improved overall speech production accuracy, positive effects in the latter group were found only for the practiced items.

Evidence indicates that developmental **stuttering** is associated with dysfunctional sensorimotor integration. Bihemispheric activation competing for control of the speech production mechanisms and atypical right-hemisphere dominance have been suggested as well (Hickok et al., 2011). More specifically, stuttering may be caused by difficulties in transitioning between sounds (Hardcastle & Tjaden, 2011), which obviously would affect coarticulation.

Studies of coarticulation in persons who stutter are inconclusive. Frisch et al. (2016) found (examining velar-vowel coarticulatory patterns) that people who stutter do not differ significantly in anticipatory coarticulation patterns from fluent speakers but that their speech stability is lower and overall variability greater, which places them at the "less skilled" end of the typical speech production range in terms of motor skill, but implies that their motor programming ability is intact. Similar conclusions were reached by Smith et al. (2010) in a study of lip aperture. They conclude that results of research into anticipatory coarticulation may have implications for intervention planning: significant differences in coarticulation patterns between fluent and stuttering output reveal higher (cognitive/linguistic) level of impairment, requiring phonologically founded treatment targeting phonological representations, whereas lack of such differences is more congruent with sensory-motor impairment that would benefit

from articulatory training, which is in line with the notion that speech motor learning is comparable to motor learning in general (Maasen et al., 2004; Donnarumma, 2017).

Acoustic analyses (expressed in terms of locus equation) in many studies of speech production in persons who stutter report atypical (steeper or shallower) or absent F_2 transitions, but at least some studies suggest normal F2 transitions in (perceptually) fluent tokens. Comparison of locus equation slopes and y-intercepts of perceptually fluent tokens of speech of children who stutter with non-stuttering controls revealed no significant differences, but F_2 transition rate was different between the two groups (for discussion, see Hardcastle & Tjaden, 2011).

According to Löfqvist (2010, p. 355) "If an articulatory pattern is to be maintained and transmitted across generations of speakers, the pattern would have to either be recoverable by auditory or audiovisual means, or follow from general principles of biomechanics and motor control." For articulators that are not visible to the naked eye (e.g., the velum or the larynx) auditory control is necessary not only for hearing and comprehension but also for learning speech production patterns. In postlingually **hearing-impaired** individuals speech production patterns/programs are maintained due to the kinesthetic and proprioceptive 'imprints' (a sort of an internal model) and somatosensory feedback. The quality and duration of these imprints depend on a number of factors, e.g., the time (childhood or adulthood) and dynamics of onset (gradually or suddenly), shape of residual hearing (favoring low or high frequency range), etc.; but in congenitally or prelingually hearing impaired individuals there is a high correlation between the degree of hearing loss and severity of speech production impairment. Smaller context effects are typically found in these subjects in comparison with normally hearing or postlingually deafened ones, which is manifested as reduced coarticulation, both anticipatory and carry-over (see Hardcastle & Tjaden, 2011, for a review).

8 Conclusions

The aim of this paper was to present some of the commonly addressed facets of coarticulation and to expose its aspects that have not received the attention they deserve. The issue of universality (biomechanics) vs. language specificity is regularly discussed in relation to carry-over and anticipatory coarticulation, respectively, although there is evidence that correlations are not exclusive. The fact that the range and direction of coarticulation are affected by a number of constraints introduces a high amount of variability in results, which makes comparisons across studies very difficult. Equally problematic for the search for coarticulation patterns are the levels at which coarticulation can be expected (labial, lingual, velar and laryngeal) and correlations among them. Related to that, models and theories that have been proposed over the years (e.g., target

undershoot, articulatory syllable, look-ahead, window, coproduction, DAC) have typically focused on only selected levels and based their assumptions on differently defined aspects of coarticulation. With the advancement of technology in the past 25 years, research into neural control of speech perception and production has progressed from speculation to actual recordings of intact central neural mechanisms at work, but reliable paradigms for studying neurophysiological bases of coarticulation have yet to be designed. Sensitive populations, such as children and individuals with various disorders (e.g., dysarthria, apraxia of speech, stuttering or hearing impairment) have so far provided inconclusive data on the nature of coarticulatory processes, apart from the general finding of greater variability than is found in the typical adult participants. Also, more research is necessary that would tie together perceptual and production results.

The issue of coarticulation is, obviously, far from being comprehensively described or defined. Approaching it from various directions: theoretical, developmental and clinical, taking into consideration its production and perception aspects, analyzing articulatory and acoustic data, using all available tools and methods (e.g., EPG, ultrasound, EMA, fMRI, locus equation, acoustic analysis), and sharing cross-linguistic data, may eventually offer converging evidence about its scope, function and control.

Acknowledgments

Work on this manuscript was carried out as part of the project Coarticulation in Croatian speech: Instrumental investigation (CROCO), IP-2016-06-5367, financed by Croatian Science Foundation. Heartfelt thanks to M. Gósy, M. Liker and anonymous reviewers for their helpful and constructive comments and suggestions.

References

Aldridge, D., Theodoros, D., Angwin, A., & Vogel, A. P. (2016). Speech outcomes in Parkinson's disease after subthalamic nucleus deep brain stimulation: A systematic review. *Parkinsonism and Related Disorders, 33,* 3-11.

Bakran, J. (1993). Kontekstualno neovisan akustički opis mjesta artikulacije okluziva: usporedna višejezična analiza. *Govor, 10*(2), 15-29.

Bakran, J., & Mildner, V. (1995). Effect of speech rate and coarticulation strategies on the locus equation determination. In *Proceedings of the XIIIth International Congress of Phonetic Sciences,* Vol. 1. (pp. 26-29), Stockholm.

Beddor, P. S., McGowan, K. B., Boland, J. E., Coetzee, A. W., & Brasher, A. (2013). The time course of perception of coarticulation. *Journal of the Acoustical Society of America, 133,* 2350-2366.

Bell-Berti, F., & Harris, K. S. (1975). Coarticulation in VCV and CVC utterances: some EMG data. *Journal of the Acoustical Society of America, 57,* suppl. 1, S70-71.

Bouchard, K. E., & Chang, E. F. (2014). Control of spoken vowel acoustics and the influence of phonetic context in human speech sensorimotor cortex. *The Journal of Neuroscience, 34*(38), 12662-12677.

Brendel, B., Hertrich, I., Erb, M., Lindner, A., Riecker, A., Grodd, W., & Ackermann, H. (2010). The contribution of mesiofrontal cortex to the preparation and execution of repetitive syllable productions: and fMRI study. *Neuroimage, 50*(3), 1219-1230.

Chang, S.-E., Kenney, M. K., Loucks, T. M. J., Poletto, C. J., & Ludlow, C. L. (2009). Common neural substrates support speech and non-speech vocal tract gestures. *Neuroimage, 47,* 314-325.

Cheng, H. Y., Murdoch, B., Goozée, J., & Scott, D. (2007). Electropalatographic assessment of tongue-to-palate contact patterns and variability in children, adolescents and adults. *Journal of Speech, Language and Hearing Research*, *50*, 375-392.

Cho, T., Kim, D., & Kim, S. (2017). Prosodically-conditioned fine-tuning of coarticulatory vowel nasalization in English. *Journal of Phonetics, 64,* 71-89.

Chomsky, N., & Halle, M. (1968). *The Sound Pattern of English.* New York: Harper and Row.

Dahan, D, Magnuson, J. S., Tanenhaus, M. K., & Hogan, E. (2001). Subcategorical mismatches and the time course of lexical access: Evidence for lexical competition. *Language and Cognitive Processes, 16,* 507-534.

Daniloff, R., & Hammarberg, R. (1973). On defining coarticulation. *Journal of Phonetics, 1,* 239-248.

Davis, C., Kleinmann, J. T., Newhart, M., Gingis, L., Pawlak, M., & Hillis, A. E. (2008). Speech and language functions that require a functioning Broca's area. *Brain and Language, 105,* 50-58.

De Smet, H. J., Baillieux, H., De Deyn, P. P., Mariën, P., & Paquier, P. (2007). The cerebellum and language: the story so far. *Folia Phoniatrica et Logopaedica, 59*(4), 165-170.

Donnarumma, F., Dindo, H., & Pezzulo, G. (2017). Sensorimotor coarticulation in the execution and recognition of intentional actions. *Frontiers in Psychology, 8,* 237. doi:10.3389/fpsyg.2017.00237.

Dronkers, N. F. (1996). A new brain region for coordinating speech articulation. *Nature, 384,* 159-161.

Farnetani, E., & Recasens, D. (2010). Coarticulation and connected speech processes. In W. J. Hardcastle, J. Laver, & F. E. Gibbon (Eds.), *The Handbook of Phonetic Sciences (2nd edition)* (pp. 316-352*).* Chichester: Wiley-Blackwell.

Fowler, C. (1980). Coarticulation and theories of extrinsic timing. *Journal of Phonetics, 8,* 113-133.

Fowler, C. & Saltzman, E. (1993). Coordination and coarticulation in speech production. *Language and Speech, 36,* 171-195.

Frisch, S. A., Maxfield, N., & Belmont, A. (2016). Anticipatory coarticulation and stability of speech in typically fluent speakers and people who stutter. *Clinical Linguistics and Phonetics, 30*(3-5), 277-291.

Goffman, L., Smith, A., Heisler, L., & Ho, M. (2008). The breadth of coarticulatory units in children and adults. *Journal of Speech, Language, and Hearing Research, 51*(6), 1424-1437.

Gordon, N. (1996). Speech, language and the cerebellum. *European Journal of Disorders of Communication, 31*(4), 359-367.

Gracco, V. L., & Löfqvist, A. (1994). Speech motor coordination and control: Evidence from lip, jaw and laryngeal movements. *The Journal of Neuroscience, 14*(11), 6585-6597.

Grigos, M. I., & Case, J. (2017). Changes in movement transitions across a practice period in childhood apraxia of speech. *Clinical Linguistics and Phonetics,* doi: 10.1080/02699206.2017.1419378.

Hardcastle, B., & Tjaden, K. (2011). Coarticulation and speech impairment. In M. J. Ball, M. R. Perkins, N. Müller, & S. Howard (Eds.), *The Handbook of Clinical Linguistics* (pp. 506-524). Blackwell Publishing Ltd.

Hickok, G., Houde, J., & Rong, F. (2011). Sensorimotor integration in speech processing: computational basis and neural organization. *Neuron, 69,* 407–422.

Hills, A. E., Work, M., Barker, P. B., Jacobs, M. A., Breese, E. L., & Maurer, K. (2004). Re-examining the brain regions crucial for orchestrating speech articulation. *Brain, 127,* 1479-1487.

Hoole, P. (1999). Coarticulatory investigations of the devoicing gesture. In W. J. Hardcastle, & N. Hewlet (Eds.), *Coarticulation: theory, data and techniques* (pp. 107-122). Cambridge: Cambridge University Press.

Horga, D., & Liker, M. (2015). *Artikulacijska fonetika: Anatomija i fiziologija izgovora* [Articulatory phonetics: Anatomy and physiology of speech production]. Zagreb: Ibis grafika.

Iskarous, K., Mooshammer, C., Hoole, P., Recasens, D., Shadle, C. H., Saltzman, E., & Whalen, D. H. (2013). The coarticulation/invariance scale: Mutual information as a measure of coarticulation resistance, motor synergy, and articulatory invariance. *Journal of the Acoustical Society of America, 134*(2), 1271-1282.

Katz, W. F. (2000). Anticipatory coarticulation and aphasia: Implications for phonetic theories. *Journal of Phonetics, 28,* 313-334.

Katz, W. F., & Bharadwaj, S. (2001). Coarticulation in fricative-vowel syllables produced by children and adults: a preliminary report. *Clinical Linguistics and Phonetics, 15,* 139-143.

Katz, W. F., Kripke, C., & Tallal, P. (1991). Anticipatory coarticulation in the speech of adults and young children: Acoustic, perceptual and video data. *Journal of Speech and Hearing Research, 34,* 1222-1232.

Keating, P. (1990). Phonetic representations in a generative grammar. *Journal of Phonetics, 18,* 321-334.

Keating, P., & Lahiri, A. (1993). Fronted velars, palatalized velars, and palatals. *Phonetica, 50,* 73-101.

Kent, R. D., Adams, S. G., & Turner, G. S. (1996). Models of speech production. In N. J. Lass (Ed.), *Principles of experimental phonetics* (pp. 3-45). St. Louis: Mosby.

Kent, R. D., & Minifie, F. D. (1977). Coarticulation in recent speech production models. *Journal of Phonetics, 5,* 115-133.

Krull, D. (1989). Second formant locus pattern and consonant-vowel coarticulation in spontaneous speech. In *Phonetic experimental research at the Institute of linguistics. PERILUS X* (pp. 43-61). Stockholm: University of Stockholm.

Leonard, M. K., Bouchard, K. E., Tang, C., & Chang, E. F. (2015). Dynamic encoding of speech sequence probability in human temporal cortex. *The Journal of Neuroscience, 35*(18), 7203-7214.

Liker, M., & Gibbon, F. (2018). Tongue-palate contact timing during /s/ and /z/ in English. *Phonetica, 75,* 110-131.

Liker, M., Mildner, V., & Šindija, B. (2007). Acoustic analysis of the speech of children with cochlear implants: a longitudinal study. *Clinical Linguistics and Phonetics, 21*(1), 1-11.

Lindblom, B. (1963). Spectrographic study of vowel reduction. *Journal of the Acoustical Society of America, 35,* 1773-1781.

Lindblom, B. (1990). Explaining phonetic variation: A sketch of the H & H theory. In W. J. Hardcastle, & A. Marchal (Eds.), *Speech Production and Speech Modeling* (pp. 403-440). Dordrecht: Kluwer.

Löfqvist, A. (1999). Interarticulator phasing, locus equations, and degree of coarticulation. *Journal of the Acoustical Society of America, 106*(4), 2022-2030.

Löfqvist, A. (2010). Theories and models of speech production. In W. J. Hardcastle, J. Laver, & F. E. Gibbon (Eds.), *The Handbook of Phonetic Sciences (2nd edition)* (pp. 353-377). Chichester: Wiley-Blackwell.

Lubker, J. F., McAllister, R. & Carlson, P. (1975). Labial co-articulation in Swedish: a preliminary report. In C. G. M. Fant (Ed.), Proceedings of the Speech Communication Seminar (pp. 55-64). Stockholm: Almqvist and Wiksell

Maasen, B., Nijland, L. & van der Muelen, S. (2001). Coarticulation within and between syllables by children with developmental apraxia of speech. *Clinical Linguistics and Phonetics, 15,* 145-150.

Maasen, B., Kent, R., Peters, H., van Lieshout, P., & Hulstijn, W. (2004). *Speech motor control in normal and disordered speech.* Oxford: Oxford University Press.

Mildner, V., & Liker, M. (2008). Fricatives, affricates and vowels in Croatian children with cochlear implants. *Clinical Linguistics and Phonetics, 22*(10-11), 845-856.

Moon, S. J., & Lindblom, B. (1994). Interaction between duration, context and speaking style in English stressed vowels. *Journal of the Acoustical Society of America, 96,* 40-55.

Mücke, D., Hermes, A., & Cho, T. (2017). Mechanisms of regulation in speech: Linguistic structure and physical control system. *Journal of Phonetics, 64,* 1-7.

Nakai, Y., Jeong, J., Brown, E. C., Rothermel, R., Kojima, K., Kambara, T., Shah, A., Mittal, S., Sood, S., & Asano, E. (2017). Three- and four-dimensional mapping of speech and language in patients with epilepsy. *Brain, 140,* 1351-1370.

Nittrouer, S., & Whalen, D. H. (1989). The perceptual effects of child-adult differences in fricative-vowel coarticulation. *Journal of the Acoustical Society of America, 86*(4), 1266-1276.

Nittrouer, S., Studdert-Kennedy, M., & Neely, S. T. (1996). How children learn to organize their speech gestures: further evidence from fricative-vowel syllables. *Journal of Speech, Language, and Hearing Research, 39*(2), 379-389.

Peeva, M. G., Guenther, F. H., Tourville, J. A., Nieto-Castanon, A., Anton, J.-L., Nazarian, B., & Alario, F.-X. (2010). Distinct representations of phonemes, syllables, and supra-syllabic sequences in the speech production network. *Neuroimage, 50,* 626-638.

Recasens, D. (1999). Lingual coarticulation. In W. J. Hardcastle, & N. Hewlet (Eds.), *Coarticulation: theory, data and techniques* (pp. 80-104). Cambridge: Cambridge University Press.

Recasens, D. (2015). The Effect of Stress and Speech Rate on Vowel Coarticulation in Catalan Vowel-Consonant-Vowel Sequences. *Journal of Speech, Language, and Hearing Research, 58*(5), 1407-1424.

Recasens, D., Pallares, M. D., & Fontdevila, J. (1997). A model of lingual coarticulation based on articulatory constraints. *Journal of the Acoustical Society of America, 102*(1), 544-561.

Riecker, A., Mathiak, K., Wildgruber, D., Erb, M., Hertrich, I., Grodd, W., & Ackermann, H. (2005). fMRI reveals two distinct cerebral networks subserving speech motor control. *Neurology, 64*(4), 700-706.

Rizzolatti, G., & Craighero, L. (2004). The mirror-neuron system. Annual Review of *Neuroscience, 27,* 169-192.

Salverda, A. P., Kleinschmidt, D., & Tanenhaus, M. K. (2014). Immediate effects of anticipatory coarticulation in spoken-word recognition. *Journal of Memory and Language, 71*(1), 145-163.

Sauvageau, V. M., Macoir, J., Langlois, M., Prud'Homme, M., Cantin, L., & Roy, J.-P. (2014). Changes in vowel articulation with subthalamic nucleus deep brain stimulation in dysarthric speakers with Parkinson's disease. *Parkinson's Disease,* 2014, Article ID 487035, doi: 10.1155/2014/487035.

Schötz, S., Frid, J., & Löfqvist, A. (2013). Development of speech motor control: lip movement variability. *Journal of the Acoustical Society of America, 133*(6), 4210-4217.

Sereno, J. A, Baum, S. R., Marean, G. C., & Lieberman, P. (1987). Acoustic analyses and perceptual data on anticipatory labial coarticulation in adults and children. *Journal of the Acoustical Society of America, 81*(2), 512-519.

Smith, A., Sadagopan, N., Walsh, B., & Weber-Fox, C. (2010). Increasing phonological complexity reveals heightened instability in inter-articulatory coordination in adults who stutter. *Journal of Fluency Disorders, 35,* 1-18.

Sussman, H. M., Hoemeke, K. A., & Ahmed, F. S. (1993). A cross-linguistic investigation of locus equations as a source of relational invariance for stop place categorization. *Journal of the Acoustical Society of America, 94,* 1256-1268.

Sussman, H. M., Marquardt, T. P., & Doyle, J. (2000). An acoustic analysis of phonemic integrity and contrastiveness in developmental apraxia of speech. *Journal of Medical Speech-Language Pathology, 8,* 301-313.

Tabain, M. (2000). Coarticulation in CV syllables: a comparison of locus equation and EPG data. *Journal of Phonetics, 28,* 137-159.

Tjaden, K. (2000). An acoustic study of coarticulation in dysarthric speakers with Parkinson disease. *Journal of Speech, Language, and Hearing Research, 43*(6), 1466-1480.

Volenec, V. (2015). Coarticulation. In J. Davis (Ed.), *Phonetics: Fundamentals, Potential Applications and Role in Communicative Disorders* (pp. 47-86). New York: Nova.

Wang, E. Q., Metman, L. V., Bakay, R. A. E., Arzbaecher, J., Bernard, B., & Corcos, D. M. (2006). Hemisphere-specific effects of subthalamic nucleus deep brain stimulation on speaking rate and articulatory accuracy of syllable repetitions in Parkinson's disease. *Journal of Medical Speech-Language Pathology, 14*(4), 323-334.

Zharkova, N. (2017). Voiceless alveolar stop coarticulation in typically developing 5-year-olds and 13-year-olds. *Clinical Linguistics and Phonetics, 31,* 503-513.

Zharkova, N., Gibbon, F. E., & Hardcastle, W. J. (2015). Quantifying lingual coarticulation using ultrasound imaging data collected with and without head stabilization. *Clinical Linguistics and Phonetics, 29,* 249-265.

Zharkova, N., Hewlett, N, & Hardcastle, W. J. (2011). Coarticulation as an indicator of speech motor control development in children: An ultrasound study. *Motor Control, 15,* 118-140.

Zharkova, N., Hewlett, N., & Hardcastle, W. J. (2012). An ultrasound study of linguistic coarticulation in /sV/ syllables produced by adults and typically developing children. *Journal of the International Phonetic Association, 42,* 193-208.

Ziegler, W., & von Cramon, D. (1986). Disturbed coarticulation in apraxia of speech: acoustic evidence. *Brain and Language, 29*(1), 34-47.

SPEECH RATE AND VOWEL QUALITY EFFECTS ON VOWEL-RELATED WORD-INITIAL IRREGULAR PHONATION IN HUNGARIAN

Alexandra MARKÓ[1,4], Andrea DEME[1,4], Márton BARTÓK[1,4],
Tekla Etelka GRÁCZI[2,4], & Tamás Gábor CSAPÓ[3,4]
[1]Department of Phonetics, Eötvös Loránd University, Budapest, Hungary,
[2]Research Institute for Linguistics, Hungarian Academy of Sciences, Budapest, Hungary, [3]Department of Telecommunication and Media Informatics, Budapest University of Technology and Economics, Budapest, Hungary,
[4]MTA–ELTE "Lendület" Lingual Articulation Research Group, Budapest, Hungary
marko.alexandra@btk.elte.hu, deme.andrea@btk.elte.hu, bartokmarton@gmail.com, graczi.tekla.etelka@nytud.mta.hu, csapot@tmit.bme.hu

Abstract
We examined utterance-initial irregular phonation as a function of vowel quality (vowel height and backness), and speech rate in Hungarian. In the analysis we distinguished two types of irregular phonation: glottalization and glottal stop. In Experiment 1, all nine Hungarian vowel qualities were analysed in pseudo words, with respect to the extent they facilitate the occurrence of irregular phonation as a function of their (i) vowel height (three levels: close, mid, open), (ii) backness using two levels in the first run (front vs. back) and three levels in the second run (front vs. central vs. back), and (iii) speech rate. In Experiment 2, four vowel qualities were analysed in real Hungarian words with respect to all the above factors (but in this analysis, only two categories were distinguished in the backness dimension). With respect to vowel height, we found that open vowels elicited more irregular phonation than mid and close vowels in both experiments. With respect to backness, in the twofold comparison (front vs. back) we found no effect in either of the experiments, while in the threefold comparison (front vs. central vs. back) we found that back vowels showed a higher ratio of irregular phonation than central and front ones in Experiment 1. The frequency of occurrence of irregular phonation was higher in fast than in slow speech in Experiment 1, and it was lower in Experiment 2 (in the latter, the confounding effect of the hiatus position was eliminated which was probably present in Experiment 1). The relative frequency of glottalization did not show an increase as a function of increased speech rate as claimed by earlier studies.

Keywords: irregular phonation, glottal stop, glottalization, vowel quality, speech rate

DOI: http://doi.org/10.18135/CAPSS.49

1 Introduction

In the present study we analyse irregular phonation in utterance-initial vowels, and we address the questions if and how the quality of the vowel and the speech rate affect its frequency of occurrence. For some aspects of these questions evidence has already been gathered in several languages, but a systematic analysis of the factors of vowel quality and the speech rate (and their interaction) has not been carried out so far. Moreover, in most cases the effect of speech rate was evaluated in a not very strictly controlled experimental design, which also raises questions with respect to the generality of conclusions drawn from these previous results. Our main questions are if vowels may elicit the occurrence of irregular phonation to a different extent as a function of their quality (with special attention paid to their height and backness features), and if speech rate affects the frequency of occurrence of irregular phonation if it is analysed in laboratory speech where the increase of speech rate is elicited in a well-controlled fashion. In addition to the frequency of occurrence of vowels realized with irregular phonation, patterns of frequency of occurrence of glottalization and glottal stops in terms of vowel quality and speech rate were also analysed to reveal the interrelations of speech rate and the type of irregular phonation the vowels are realized with.

1.1 Irregular phonation

Modal voice is defined in the literature as quasi-periodic vibration of the glottal folds (e.g., Gósy, 2004), and is considered to be the most common type of phonation. However, in some cases, voice production may depart from this typical pattern, and phonation may become irregular. Irregular phonation is used as an umbrella term in the literature, covering several types of irregularity in vocal fold vibration. Besides *irregular phonation,* other terms like *laryngealization, glottalization, creaky voice,* etc. are also used, and in several cases they refer to only more or less similar realizations of irregularity in the voice source. Based on their formal characteristics, some authors use more accurate definitions for the subtypes of irregular phonation, (e.g., Batliner et al., 1993; Dilley et al., 1996), while in several studies the concept of irregularity is introduced in a more intuitive manner. Considering the terminological variability, it is crucial that we clarify our use of terms in the present work. We refer to irregularity in the voice source in general using the term *irregular phonation.* In the present study, we investigate two easily distinguishable types of irregular phonation. For one of these phenomena we apply the term *glottalization* (covering several possible subtypes) to refer to cases where irregularity can be observed as consecutive periods in voicing differing evidently in terms of duration, amplitude, or both. The second phenomenon is a single glottal gesture which we refer to as *glottal stop* (Figure 1).

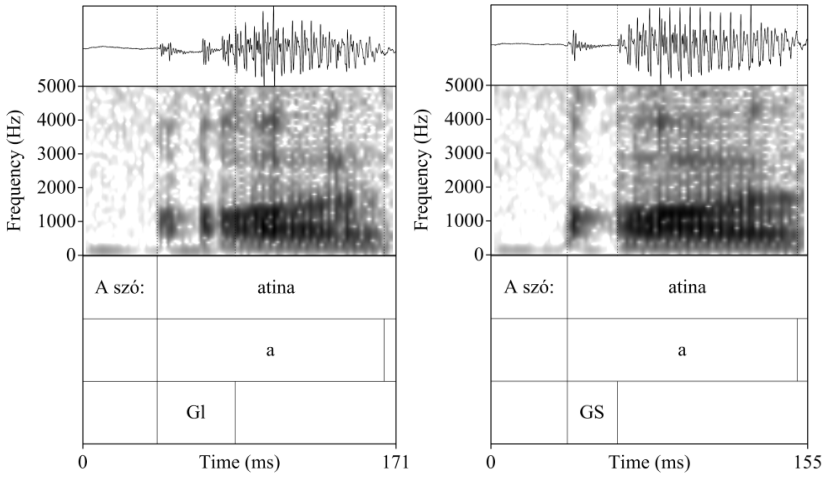


Figure 1.
Examples of glottalization (left) and glottal stop (right)

Irregular phonation serves prosodic functions in a number of typologically unrelated languages, e.g., American English (Dilley et al., 1996), Czech, Spanish (Bissiri et al., 2011), German, Polish (Kohler, 2001; Malisz et al., 2013), Hungarian (Markó, 2013) and others. The occurrence of irregularity in the voice source may be influenced by several factors (see some of these below), and it shows high inter- and intraspeaker variability (e.g., Dilley et al., 1996; Redi & Shattuck-Hufnagel, 2001). It was also shown that the less speakers glottalize, the more probable it is that they do so in a phrase-, word- or vowel-to-vowel boundary (hiatus) position (Markó, 2013).

A number of studies have found irregular phonation to predominate among male speakers in several speech communities (e.g., Stuart-Smith (1999) in Glasgow; Esling (1978) in Edinburgh; Henton and Bladon (1988) for speakers of RP and 'Modified Northern' English). Nevertheless, despite strong associations between irregular phonation and the male gender, the opposite tendency is also documented in the literature. For example, irregular phonation was found to be prevalent in college-aged women in Virginia (Lefkowitz, 2007, cited by Podesva, 2013), and young Californian women also use it significantly more often than their male counterparts (Yuasa, 2010). Podesva (2013) found similar tendencies independently of age and race in the Washington, DC, Metropolitan Area. In Hungarian, irregular phonation was found to be more frequent in young and middle-aged females' speech than in male speakers of the same age groups (see Markó, 2013).

Kohler (2001, pp. 282–285) defined four types of irregular phonation (which he generally labelled as *glottalization* covering "the glottal stop and any devia-

tion from canonical modal voice") as follows. (1) Vowel-related glottalization phenomena which signal the boundaries of words or morphemes. (2) Plosive-related glottalization phenomena which occur as reinforcement or even replacement of plosives. (3) Syllable-related glottalization phenomena which characterize syllable types along a scale from a glottal stop to glottalization (e.g., Danish stød). (4) Utterance-related glottalization phenomena which comprise (i) phrase-final relaxation of phonation, and (ii) truncation glottalization, i.e., utterance-internal tensing of phonation at utterance breaks.

Initial irregular phonation in vowels (a specific case of type (1) above) was analyzed in several studies. Malisz et al. (2013) examined the conditioning effect of speech style (speech vs. dialogue), presence of prominence, phrasal position (initial vs. medial), speech rate, word type, preceding segment, and following vowel height on the frequency of occurrence of word-initial (and vowel-related) glottalization in Polish and German. They concluded – among other points – that vowels bearing prominence were more frequently marked glottally (in both languages); faster rates reduced glottal marking in general, but especially the number of glottal stops; and that faster rates increased the relative frequency of the occurrence of glottalization. They also found that low vowels were more frequently glottalized in both languages than non-low vowels; however, it must also be noted that the factors of speech rate and vowel quality were not systematically varied in this study.

Lancia and Grawunder (2014) used pseudo-words to facilitate initial irregular phonation in vowels, and to analyze the conditioning factors of vowel height (high vs. low: /i/ vs. /a/), the presence of stress, and the place of articulation of the preceding consonants. They concluded that retracted tongue (i.e., low/back tongue position) favors the production of irregular phonation (particularly strongly in unstressed syllables).

In Hungarian a systematic analysis of the effect of speech rate and vowel quality in initial irregular phonation in vowels has not been carried out so far; in addition, to the authors' knowledge, a study considering the interaction of these factors, or the separate analysis of the effect of such vowel features as vowel height and backness is also nonexistent for any other languages either. Moreover, in earlier studies investigating vowel-related voice source irregularity in Hungarian (e.g., Markó, 2013), glottalization and glottal stops were not treated as separate categories; as a result, no data is available on the relative frequency of these two types either.

1.2 Speech rate

There are a number of measures that are used to parameterize speech rate in timing studies from average syllable duration (ASD) to words/minute (see an overview in Fletcher, 2010). If we want to compare tempo characteristics of

segmentally identical sentences, it is also a possibility to simply compare total sentence durations (see e.g., Smith, 2010).

Global and local speech rate measures can be differentiated; however, the use of these terms is rather confusing in the literature. In the present study we refer to speech rate as a global characteristics of the utterance, where micro-temporal variation (like, e.g., phrase-final lengthening) is not considered in more detail.

1.3 The Hungarian vowel inventory

The Hungarian vowel inventory includes 14 vowels (see Figure 2), which are paired in the dimension of quantity resulting in 7 short-long phonological pairs. However, the members of two short-long pairs (/ɛ/ and /eː/; /ɒ/ and /aː/) differ in their phonetic characteristics as well; therefore, from a phonetic point of view, 9 vowel qualities can be differentiated in Hungarian.

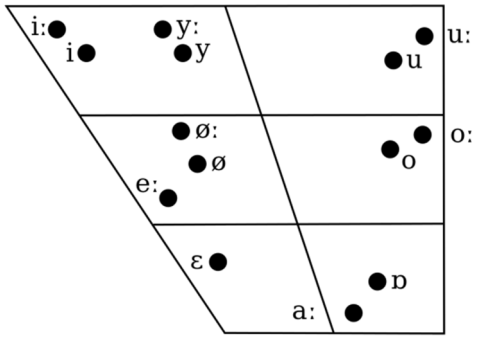


Figure 2.
The Hungarian vowel inventory (after Szende, 1994: 92)[1]

According to the traditional view in Hungarian phonetics (see, e.g., the textbooks of Kassai, 1998; Gósy, 2004), vowels are differentiated on the vowel height dimension as follows: /i y u/ are considered as close vowels, /eː ø o/ are categorized as close-mid, /ɛ/ is considered as open-mid, and /aː/ is considered as an open vowel. The short phonological counterpart of /aː/ in this view is considered to be the open-mid /ɔ/, while others (e.g., Mády, 2008) define the vowel at hand as an open /ɒ/. (Note that in the present paper, we adhere to the latter notation and analysis.)

[1]SVG version of IPA vowel chart for Hungarian, showing the "short a" as a rounded vowel in contrast to Szende (1994).
https://commons.wikimedia.org/wiki/File:Hungarian_vowel_chart_with_rounded_short_a.svg

According to the traditional view again, with respect to backness, /i y eː ø ɛ/ are considered as front vowels, while /u o ɒ aː/ are characterized as back vowels. It should also be noted, however, that the status of the vowel /aː/ is ambiguous: while it is uniformly transcribed with the IPA symbol of a front vowel, it is generally classified as a back vowel (based on its morpho-phonological behavior, namely its participation in the Hungarian vowel harmony) both by the phonological (e.g., Siptár & Törkenczy, 2007) and the phonetic literature.

Although the characteristics of the vowel system on the height dimension are more or less generally agreed on, with respect to the backness distinction there is another competing view on the vowel system which was introduced by Bolla (1995). On the basis of the articulatory (X-ray) analysis of Hungarian vowels, Bolla (1995) claimed that /i eː ɛ/ are front vowels, /y ø aː/ are central vowels, while /u o ɒ/ are back vowels.

Since the present study focuses on the question if vowels may elicit the occurrence of irregular phonation to a different extent as a function of their quality with special attention paid to their backness and vowel height features, we will introduce both the twofold and the threefold analysis of the backness feature in the analysis of our data, where applicable, to get a deeper insight into the question at hand (see Experiment 1, and sections 2.2.2 and 2.2.3). Additionally, through the application of the threefold opposition, we also hope to resolve the bias inherent in the twofold analysis due to the ambiguous status of /aː/ (namely, that it is a central vowel but categorized as a back vowel due to its phonological behaviour).

1.4 Aims and hypotheses

In the present study we investigated the effect of vowel height and backness and speech rate in two experiments, using phonetically balanced speech materials, which were also carefully controlled with respect to speech rate. In the analysis, two types of irregular phonation were taken into account: glottalization and glottal stop. In Experiment 1 we recorded pseudo-words in order to analyse all of the 9 different vowel qualities of Hungarian (irrespective of quantity, i.e., /i y u eː ø o ɛ ɒ aː/, see Figure 2) in the same context. In Experiment 2, we used the same design with real words, and analysed 4 vowel qualities /i o ɛ ɒ/.

Based on previous results for other languages, we addressed the following questions. Is irregular phonation more frequent in word-initial vowels in Hungarian if (i) the speech rate is slow (as opposed to fast); (ii) the vowel is back (as opposed to front); (iii) the vowel is open (as opposed to close or close-mid)? (iv) Do glottal stops occur less frequently in fast speech, while the relative amount of glottalization increases?

We hypothesized that the frequency of occurrence of irregular phonation in general is higher in slow speech than in fast speech. Furthermore, we assumed

that back vowels elicit irregular phonation in a higher ratio than front ones both in slow and fast speech, and that open vowels favour irregular phonation more than non-open (close or close-mid) ones due to tongue retraction associated with the back and open articulatory positions. Finally, we also assumed that faster speech rate reduces the number of glottal stops, but increases the relative frequency of glottalization.

2 Experiment 1

2.1 Material and method

2.1.1 Material

The test material consisted of trisyllabic pseudo-words which fit into the phonotactic patterns of the Hungarian language. Each of the different Hungarian vowel qualities (/i y u eː ø o ɛ ɒ aː/, see Figure 2) appeared as the first syllable of the construction V*tina,* where both word-stress (given that word stress is fixed on the first syllable in Hungarian) and pitch accent were expected in all cases. These nonsense target words were embedded in the following phrase: *A szó:* [target word] 'The word is: [target word]'.

2.1.2 Experimental design

The stimuli were presented to the speakers on a computer screen. Each trial consisted of two display screens: first the introductory part *(A szó:)* was shown to the participant, then the target word was displayed (see Figure 3). The participants' task was to read aloud the target word, but not the introductory part.

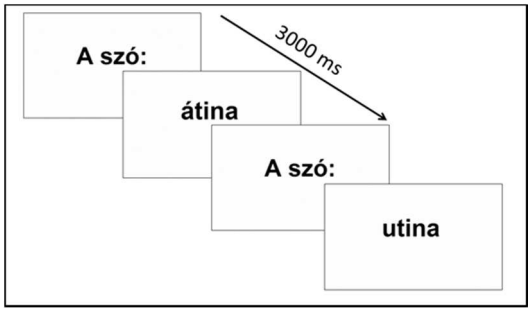


Figure 3.
Two consecutive trials with the four corresponding display screens timed 1500 ms each (resulting in 3000 ms for one trial)

In order to elicit speech rate differences between the conditions, the timing of the display screens was manipulated. In the "slow" speech condition, each display screen appeared for 1500 ms resulting in 3000 ms for one trial in total (including the introductory part and the target word). In the "fast" speech condition the timer was set to 300 ms, resulting in 600 ms for one trial in total.

(During the recordings several other timer settings were also applied, and the setting to serve as the "fast" condition was selected posterior to the recordings on the basis of the speakers' ability to produce the items properly, i.e., separately and without errors.) As the timing of the introductory part reflected also the timing of the target word, it enabled the speakers to prepare for the production of the latter.

The trials were ordered into blocks: within each block all nine different target words occurred in a randomized order once, and these blocks were repeated 5 times consecutively for each ("slow" and "fast") condition. First the "slow" condition, then the "fast" condition was recorded in the case of every participant. 9 vowel qualities/target words × 5 repetitions × 2 speech rate conditions, i.e., 90 vowels per speaker were recorded. The recordings were made in a sound-treated booth, using a tie-clip omnidirectional condenser microphone and an external soundcard.

2.1.3 Participants

As previous results revealed that Hungarian female speakers tend to produce irregular phonation more frequently than male speakers (see e.g., Markó, 2013); in the present study only female speakers were included. All 14 of them were university students, and native speakers of Hungarian, who reported no hearing or speech deficits. In order to ensure that the "slow" and the "fast" conditions differentiate properly (i.e., they may be differentiated by a conceptually sound value), a threshold for the speech rate difference was introduced (see below, in 2.1.5).

This threshold was not exceeded by the data in the case of 4 participants, thus finally in the main analysis 10 speakers' material was involved. The speakers' age ranged between 23 and 28 years, with a mean of 25 years.

2.1.4 Annotation

The target word, the word-initial vowel, and the irregular phonation at the beginning of the word-initial vowel were labelled manually in Praat (Boersma & Weenink, 2016).

The vowel qualities were identified automatically (on the basis of the stimulus order), and then checked by the annotators (two of the authors of the present paper) auditorily. In the case of mispronunciation or any other errors involving the production of the vowel of interest, the vowel was excluded from the material. Vowel boundaries were defined on the basis of the F_2 trajectory.

The labeling of the irregular phonation was performed in accordance with the methodology proposed by previous studies (e.g., Dilley et al., 1996; Bőhm & Ujváry, 2008) in which visual (waveform and spectrogram) and auditive information was combined. A given vowel was labelled as irregularly phonated if (i) its first consecutive periods differed evidently in terms of duration, amplitude, or both (these cases were marked as glottalization, Figure 1, left), or

if (ii) one (or more) glottal stop(s) was/were observed at the beginning of the vowel (these cases were marked as glottal stop, Figure 1, right).

We analysed the ratio of vowels produced with irregular phonation with respect to vowel quality, vowel height and backness, and speech rate. We also determined the ratio of glottalized vowel occurrences to the total number of vowels realized with irregular phonation, and compared that to the ratio of vowels realized with glottal stops in the two speech rate conditions.

2.1.5 Control of speech rate

Given that one of the aims of the present study was to compare "slow" and "fast" speech with respect to the frequency of occurrence of initial irregular phonation in vowels, it was inevitable to properly control for the difference of speech rate between these two conditions. To achieve this goal, first we manipulated the timing of the display screens (see above); however, according to the authors' perceptual judgement, this did not necessarily lead to considerable differences between the speech rates of the two analysed conditions speakerwise. Therefore, we decided to set a perceptually motivated threshold for the speech rate differences.

In the case of Hungarian, the just noticeable difference (JND) for speech tempo has not been studied so far; however, there are JND data for other languages which may be taken as a reasonable reference for Hungarian as well. For instance, for Dutch speech fragments Quené (2007) found 5% JND for artificially increased/decreased speech rate differences, while he also noted that this value may be an overestimation and that in the case of everyday communicative situations, the JND is probably lower.

As in our case the number of the phonemes per item was constant, to calculate speech rate differences, not the speech sound per duration values, but only the word durations were measured and compared in the two speech rate conditions. The five repetitions of each word were analysed and averaged for this comparison. The mean durations (± one SD) in "slow" and "fast" conditions for those 10 speakers who differentiated their speech rates by more than 5% on average can be seen in Figure 4. The duration difference between the two conditions ranged between 5% and 20% speakerwise, and the mean of the differences was 11±5.3%. The average item duration was 535±54 ms in the "slow" condition, and 483±52 ms in the "fast" condition (for all of the 10 speakers pooled).

2.1.6 Statistical analyses

Three 2-way repeated measures ANOVAs were performed with the factors (i) *vowel quality* and *speech rate,* (ii) *vowel height, backness,* and *speech rate,* and (iii) *vowel height, backness,* and *speech rate,* but in the latter case the feature backness was redefined to contain three levels instead of two, on the basis of the analysis of Bolla (1995) (see 1.2 and 2.2.3 for further details). The confidence level was set to 95% in each case.

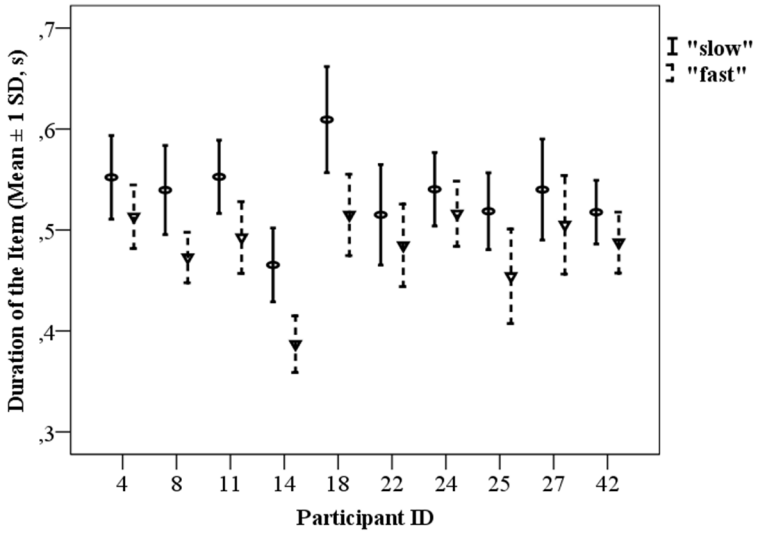


Figure 4.
The duration of items as a function of speech rate condition, speakerwise (mean ± SD)

2.2 Results

2.2.1 Vowel-initial irregular phonation as a function of vowel quality and speech rate

The ratio of vowels produced with irregular phonation (pooled over speakers) as a function of *vowel quality* and *speech rate* is presented in Figure 5.

In the "slow" condition the ratio of initial irregular phonation in vowels was 74.0±16.9% on average, while in the "fast" condition 79.5±7.2%. A repeated measures factorial ANOVA showed significant interaction of the vowel quality and the speech rate factors $F(8, 72) = 2.84$, $p = 0.008$, revealing that increased speech rate affects the frequency of occurrence of irregular phonation in the vowels differently. These differences will be further explored in the analysis by vowel features below.

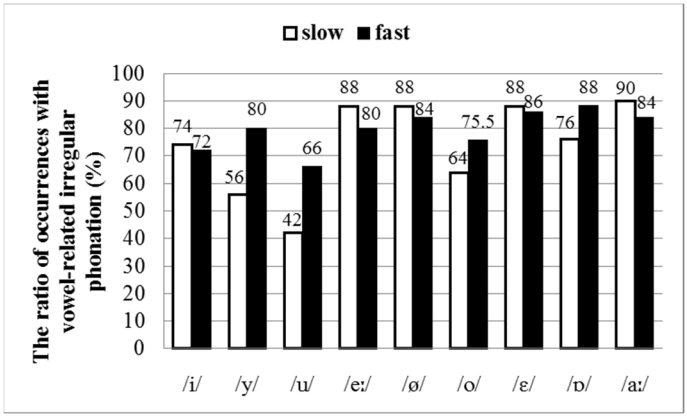


Figure 5.
The ratio of vowels produced with any kind of irregular phonation as a function of vowel quality and speech rate

2.2.2 Initial irregular phonation as a function of vowel height, backness (two levels), and speech rate

In the present analysis, on the vowel height dimension /i y u/ were considered as close vowels, /eː ø o/ were considered as mid vowels, and /ɛ ɒ aː/ were considered as open vowels. Note that this is a simplification of the traditional view of Hungarian phonetics (see 1.2), but we argue that it is highly tenable on the basis of the figure of Szende (1994), which represents the Hungarian vowel inventory (see Figure 2). Additionally, this simplification also balances the conditions numerically on a scientifically sound basis, which thus corrects the bias introduced by the imbalance of the four-level analysis.

In the first model, again on the basis of the traditional phonetic and phonological classification of Hungarian vowels, we considered /i y eː ø ɛ/ as front vowels, while /u o ɒ aː/ were characterized as back vowels (see e.g., Gósy, 2004; Siptár & Törkenczy, 2007).

The ratio of vowels produced with irregular phonation as a function of vowel height and speech rate is shown in Figure 6, while the ratio of vowels produced with irregular phonation as a function of backness (consisting of the two levels, back and front) and speech rate is shown in Figure 7.

With respect to the frequency of occurrence of initial irregular phonation in vowels, significant interaction effect of *vowel height* and *speech rate* was found ($F(2, 18) = 5.20$, $p < 0.05$), reflecting that speech rate affected vowels differently according to their height. The pairwise comparisons revealed that in the "slow" speech condition, the mid and the close vowels ($p = 0.002$), and the open and the close vowels ($p < 0.001$) differed significantly, while in the "fast" speech condition, the open and the close vowels ($p = 0.02$) were differentiated with

respect to the frequency of irregular phonation. However, we found no effect of backness (Figure 7).

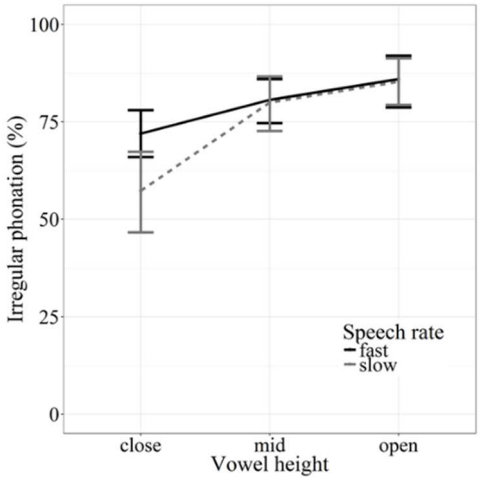


Figure 6.
The ratio of vowels produced with any kind of irregular phonation as a function of vowel height and speech rate (mean + 95% CI)

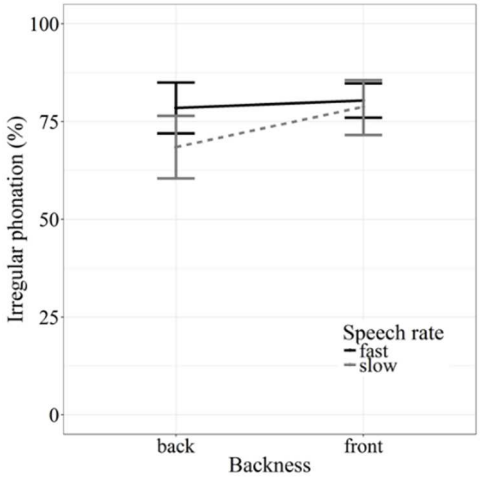


Figure 7.
The ratio of vowels produced with any kind of irregular phonation as a function of backness (two levels) and speech rate (mean + 95% CI)

2.2.3 Initial irregular phonation as a function of vowel height, backness (three levels) and speech rate

As mentioned above, another possible analysis of the backness feature in Hungarian vowels is to treat them in a threefold contrast, and differentiate back /u o ɒ/, central /y ø aː/, and front /i eː ɛ/ vowels (see Bolla, 1995, p. 211). Since in the present paper we do not intend to decide in favor of either of the possible analyses, but we acknowledge the more fine-grained nature of the latter one, we opted for testing this categorization as well, and see if this may reveal any trends that remained hidden in the model using the more traditional approach.

The ratio of vowels produced with irregular phonation as a function of backness (with three levels) and speech rate is shown in Figure 8. According to an ANOVA, the factors *vowel height, backness,* and *speech rate* display two interaction effects on the ratio of vowels realized with irregular phonation: *speech rate* interacts with *vowel height* ($F(2, 18) = 3.97$, $p < 0.05$), and *speech rate* interacts with *backness* ($F(2, 18) = 6.04$, $p < 0.01$), as well. Obviously, the interaction of *speech rate* with *vowel height* is the same effect we found while fitting the previous model. The interaction of *speech rate* and *backness* is, however, a new finding that emerged from the recategorization of the data. According to the pairwise comparisons, this interaction is due to the fact that there is a significant difference between the "slow" and "fast" conditions in the case of back vowels ($p < 0.05$), but we do not see this differentiation of the two tempo conditions in the central or in the front vowels. What is more, as opposed to the trend seen in back and central vowels, in the case of front vowels, the "slow" condition seems to have elicited more vowels with irregular phonation.

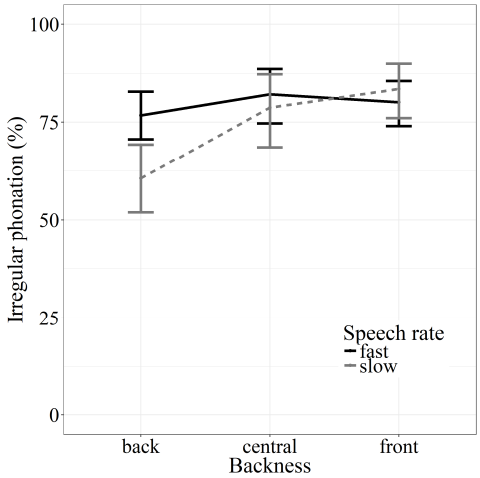


Figure 8.
The ratio of vowels produced with any kind of irregular phonation as a function of backness (three levels) and speech rate (mean + 95% CI)

2.2.4 Ratio of glottalization and glottal stops as a function of vowel quality and speech rate

Relative to the number of all irregular occurrences both in the "slow" and the "fast" conditions, the ratio of glottalization (55.9±9.2% and 66.3±9.2% of all irregular occurrences, respectively) exceeded the ratio of glottal stops (44.1±9.2% and 33.7±9.2% of all irregular occurrences, respectively) in general. The only exception we found was the vowel /u/, in which the ratio of glottal stops (66.7%) was well above the ratio of glottalization (33.3%) in the "slow" condition.

The ratio of glottalization in the "fast" condition exceeded that of the "slow" condition in all vowels but /y/. The ratio of glottalization and glottal stops were close to equal in the case of /y/ in both the "slow" (53.6 vs. 46.%) and the "fast" (52.5 vs. 47.5%) condition. The ratio of occurrences of glottal stops and glottalization as a function of vowel quality and speech rate are presented in Figure 9.

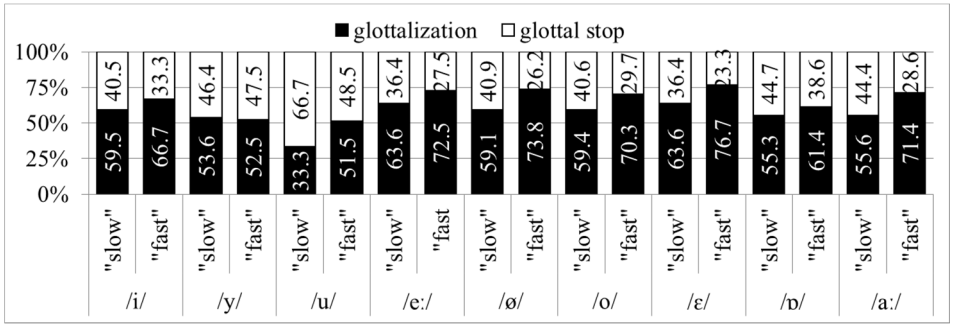


Figure 9.
The ratio of the two types of irregular phonation (relative to all irregular occurrences) as a function of vowel quality and speech rate

In the case of the close vowels, the ratio of glottalization was 51.1% in the "slow" condition and 56.9% in the "fast" condition, relative to the number of all irregular occurrences. This means that in the close vowels the ratio of glottalization and glottal stops were close to equal. The mid vowels showed the highest ratios of glottalization: 60.8% and 72.3% (in the "slow" and the "fast" conditions, respectively); while in the open vowels, 58.3% and 69.8% (in the "slow" and the "fast" conditions, respectively) of glottalization was measured.

With respect to the backness of the vowels, in the twofold comparison, the front vowels showed 60.4% of glottalization in the "slow" condition, and 68.7% in the "fast" condition, while in the back vowels these ratios were 52.9% and 64.1% in the "slow" and the "fast" conditions, respectively. In the threefold comparison, the front vowels showed glottalization in 62.4% in the "slow" condition, and in 72.3% in the "fast" condition of all cases labelled as realized with irregular phonation, the central vowels showed glottalization in 57.9% and

in 63.6% in the "slow" and in the "fast" conditions, respectively, and the back vowels showed glottalization in 50.0% in the "slow" and 64.2% in the "fast" condition. To summarize, the ratio of glottalization in the "fast" condition exceeded that of the "slow" condition in all vowel groups.

2.3 Discussion

The analysis of pseudo-words in two different speech rate conditions resulted in an unexpected difference between "slow" and "fast" speech, namely that speakers produced a higher frequency of occurrences of initial irregular phonation in vowels in "fast" speech than in "slow" speech. The statistical analysis showed a significant interaction effect of *vowel quality* and *speech rate,* and a further test showed an interaction effect of *vowel height* and *speech rate* on the frequency of occurrence of irregular phonation, revealing that vowels were affected differently by increased speech rate according to their height feature: close vowels were more susceptible to irregular phonation than mid or open vowels. When we analyzed the vowels in a finer three-fold backness contrast, the data also showed a significant *backness* × *speech rate* interaction, revealing that the effect of speech rate increased the frequency of occurrence of irregular phonation in back vowels the most, while central and front vowels did not show a clear and consistent effect.

The ratio of glottal stops (relative to the number of all irregular occurrences) was the highest in the case of close vowels, which can be interpreted as follows: the less probably irregular phonation occurs, the more likely it appears as a glottal stop.

The above mentioned unexpected results raise the question if the experimental design and the material analysed were adequate for the study. The JND-based method and the JND value (borrowed from results for Dutch, see Quené, 2007) we used to objectively define and distinguish (minimally two) speech rate conditions appeared to be suitable and appropriate for our purposes, that is, to reasonably designate groups of speech tempi which may be regarded as different groups (based on their objectively measured speech rate values). However, the criterion for the inclusion of speakers in the analysis may have led to some undesired artefacts in the results. As we defined the threshold of tempo differences on the basis of the averaged values of one speaker, we may have included several tokens of the target items in the analysis which were actually not distinguished by the previously defined threshold. Therefore, we may also have introduced some noise in the data.

As far as the target word is considered in which the target vowels were embedded, other types of complications arose. Since the V*tina* construction starts and ends with a vowel, in its consecutive production, a hiatus position occurs. The hiatus position, however, is not irrelevant to the production of

irregular phonation, as it may elicit a higher frequency of occurrence of it (by which the speakers try to avoid this disallowed phonotactic phenomenon; see, e.g., Markó, 2013). Moreover, we also suspect that this effect is not even constant across the conditions, as in slower speech the speakers are more likely to insert longer pauses between the consecutive target words, thus the effect of the hiatus may be greater in fast speech. On this basis, we replicated the experiment with several modifications. First, in Experiment 2 we used real Hungarian words instead of nonsense words. And secondly, these target words were selected to be adequate to retest the effect of the backness and height features of the vowels, while they were also sufficient to avoid the risks of the confounding effect of hiatus, since they ended with a consonant.

3 Experiment 2

3.1 Material and method

3.1.1 Material and experimental design

In Experiment 2, the test material consisted of disyllabic Hungarian pronominal adverbs: *innen* /inːɛn/ 'from here'; *onnan* /onːɒn/ 'from there'; *ennek* /ɛnːɛk/ 'for this'; *annak* /ɒnːɒk/ 'for that'. These adverbs start with four different vowel qualities which vary both in the vowel height (close /i/ vs. mid /o/ vs. open /ɛ ɒ/) and the backness (back /o ɒ/ vs. front /i ɛ/) features (while backness also co-varies with lip spreading) (see Figure 2). (It is important to note here that in Experiment 2, by the introduction of /ɛ/ as an open vowel we used a "simplified" feature set along the vowel height dimension again, to which the system described by Szende (1994) and the similar "acoustic openness" of /ɛ/ and /ɒ/ provided the basis.)

Similarly to Experiment 1, the target words were embedded in a carrier phrase. The phrase used here was the following: *Mondd:* [target word] *kell.* 'Say: [target word] needed'. All of the target vowels were positioned word-initially, thus they bore sentential accent on the first syllable (given that word stress is fixed on the first syllable in Hungarian).

Also similarly to Experiment 1, the stimuli were presented on a computer screen. Each trial consisted of two display screens again: first the introductory part *(Mondd:)* was shown to the participant, then the target item (target word + *kell)* was displayed (refer to Figure 3). The participants' task was to read aloud the target item, but not the introductory part.

In order to elicit speech rate differences between the conditions, the timing of the display screens was manipulated again. However, as the carrier phrase used in Experiment 2 was longer than the one used in Experiment 1, we used a slightly longer timer setting in the "fast" condition than previously. As a result, in the "slow" speech condition, each display screen appeared for 1500 ms

resulting in 3000 ms for one trial in total (including the introductory part and the "target word + *kell"* construction) just as in Experiment 1. However, in the "fast" speech condition the timer was set to 500 ms, resulting in 1000 ms for one trial in total. The experimental procedure was the same as in Experiment 1 in other respects. In Experiment 2, 4 vowel qualities/target words × 5 repetitions × 2 speech rate conditions, i.e., 40 vowels per speaker were recorded.

3.1.2 Participants and the control of speech rate

For Experiment 2 only female speakers were recruited; all 33 of them were university students, and native speakers of Hungarian, who reported no hearing or speech deficits. In order to ensure that the "slow" and the "fast" conditions differentiate properly (i.e., they may be differentiated by a conceptually sound value), the JND-based threshold (introduced in Experiment 1) was used again to define speech rate differences. For this purpose, we measured and compared the target item durations both in "slow" and "fast" conditions speakerwise just as in Experiment 1. However, in the present experiment, we decided to apply a higher threshold of 10% to account for the expected reduction in duration variability, since in this study, five repetitions of each item were analyzed and averaged. With this higher threshold we intended to establish a larger gap between the mean durations of the "slow" and the "fast" realizations, that is, to reduce the overlap between the conditions by eliminating most of the vowel realizations which were extremely long in the "fast", or extremely short in the "slow" conditions. We opted for the application of the stricter threshold due to the fact that in Experiment 2 we managed to recruit a higher number of participants than in Experiment 1, so that after the exclusion of speakers whose speech samples were not differentiated by the threshold, the amount of the data was still sufficient for analysis. The threshold was exceeded by the data in the case of 18 participants, thus finally in the main analysis 18 speakers' material was involved (and 15 participants' data were excluded). The speakers' age ranged between 19 and 34 years, with a mean of 24.9 years. The duration difference between the two conditions ranged between 9.6% and 28.5% speakerwise, and the mean of differences was 16.8±5.5% (Figure 10). The overall item duration was 852±107 ms in the "slow" condition, and 705±69 ms in the "fast" condition (for all 18 speakers). (We should also note that in some cases, the participants might have inserted a short pause at the word boundary of *innen/onnan/ennek/annak # kell.* However, as this pause cannot be differentiated reliably from the occlusion phase of the second plosive, we did not identify these possible pauses and regarded the total item duration as item duration in the data.)

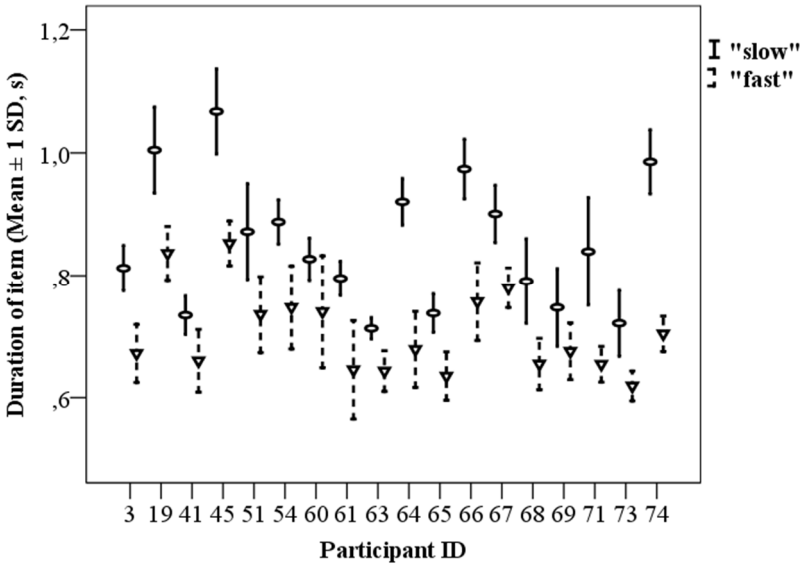


Figure 10.
The duration of items (target word + *kell)* as a function of speech rate conditions, speakerwise (mean ± SD)

3.1.3 Annotation and analysis

In the "slow" condition, 359 vowels were analyzed (as one speaker mispronounced one target word), while in the "fast" condition the number of the analyzed vowels was 360.

The "target word + *kell"* construction, the word-initial vowel, and the irregular phonation at the beginning of the word-initial vowel were labelled manually in Praat (Boersma & Weenink, 2016). The labeling was performed similarly to Experiment 1, and we used the categories of "glottalization" and "glottal stop" again.

We analysed the ratio of vowels produced with irregular phonation with respect to vowel quality, vowel height, backness, and speech rate. As previously, we determined and compared the ratio of glottalized vowel occurrences to the number of all vowels realized with irregular phonation again, and compared it to the ratio of glottal stops in the two speech rate conditions.

Three 2-way repeated measures ANOVAs were performed with the factors (i) *vowel quality* and *speech rate,* (ii) *vowel height* and *speech rate,* and (iii) *backness* and *speech rate* at a confidence level set to 95%.

3.2 Results

3.2.1 Initial irregular phonation as a function of vowel quality and speech rate

The ratio of vowels produced with irregular phonation (pooled over speakers) as a function of vowel quality and speech rate is presented in Figure 11.

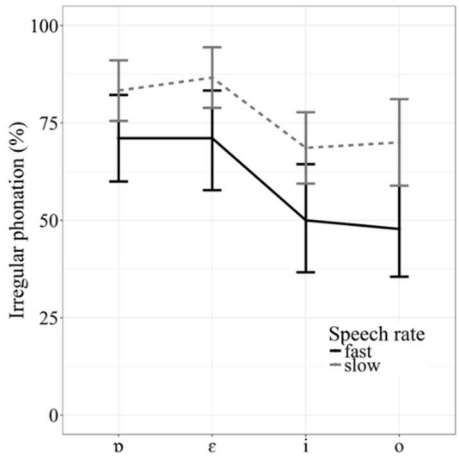


Figure 11.
The ratio of vowels produced with any kind of irregular phonation as a function of vowel quality and speech rate (mean + 95% CI)

In the "slow" condition, the ratio of initial irregular phonation in the front close /i/ was 68.6±20.7%, while in the "fast" condition it was 50.0±30.1% of all cases. In the case of the front open vowel /ɛ/ these ratios were 86.7±18.1% in the "slow" and 71.1±29.3% in the "fast" conditions. The back mid vowel /o/ was produced with irregular phonation in 70.0±24.0% of all cases in the "slow" and in 47.8±25.8% of all cases in the "fast" condition. Finally, the back open /ɒ/ showed 83.3±18.5% of irregular occurrences in the "slow", and 71.1±24.9% in the "fast" condition.

The ANOVA showed significant main effects of both *speech rate* ($F(1, 17) = 17.38$, $p < 0.001$) and *vowel quality* ($F(3, 51) = 10.56$, $p < 0.001$), but the interaction of these factors turned out to be non-significant.

3.2.2 Initial irregular phonation as a function of backness and speech rate

The ratio of vowels produced with irregular phonation as a function of backness and speech rate is shown in Figure 12.

The back /o/ and /ɒ/ and the front /ɛ/ and /i/ vowels were produced with irregular phonation in a similar ratio both in the "slow" and the "fast" conditions. In the "slow" condition front vowels showed 77.6±21.2% ratio of irregular occurrences, while back vowels showed 76.7±22.1% of all cases. In the

"fast" condition the ratios were 60.6±31.2% and 59.4±27.7%, respectively. Regarding this comparison, statistical analysis showed a significant difference between the *speech rate* conditions ($F(1, 68) = 14.58$, $p < 0.001$), but not between the front and back vowel groups. We must note here that as in the present analysis no central vowels are involved, we could not apply the three-fold analysis of vowel backness we presented in Experiment 1.

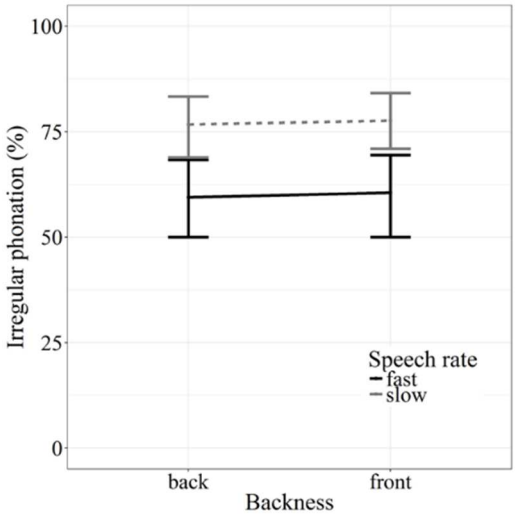


Figure 12.
The ratio of vowels produced with any kind of irregular phonation as a function of backness and speech rate (mean + 95% CI)

3.2.3 Initial irregular phonation as a function of vowel height and speech rate

The ratios of vowel realizations with irregular phonation in terms of vowel height and speech rate are shown in Figure 13. The close vowel /i/ was produced with irregular phonation in 68.6±20.7% of all cases in the "slow" and in 50.0±30.1% of all cases in the "fast" condition. The mid vowel /o/ showed initial irregular phonation in 70.0±24.0% of all cases in the "slow" and in 47.8±25.8% in the "fast" condition. The open vowels /ɒ/ and /ɛ/ were produced with irregular phonation at the highest ratio: in 85.0±18.1% in the "slow" and in 71.1±26.8% in the "fast" condition.

According to the ANOVA, there was no significant interaction between *speech rate* and *vowel height* but both factors had a significant main effect (*vowel height:* $F(2, 34) = 13.20, p < 0.001$; *speech rate:* $F(2, 17) = 17.28, p < 0.001$).

3.2.4 Ratio of glottalization and glottal stops as a function of vowel quality and speech rate

The ratios of occurrence of glottal stops and glottalization as a function of vowel quality and speech rate are presented in Figure 14. Relative to the number of all irregular occurrences both in the "slow" and the "fast" conditions, the ratio of glottalization (54.9% and 56.6.0% of all irregular occurrences, respectively) exceeded the ratio of glottal stops (45.1% and 43.4% of all irregular occurrences, respectively) as well.

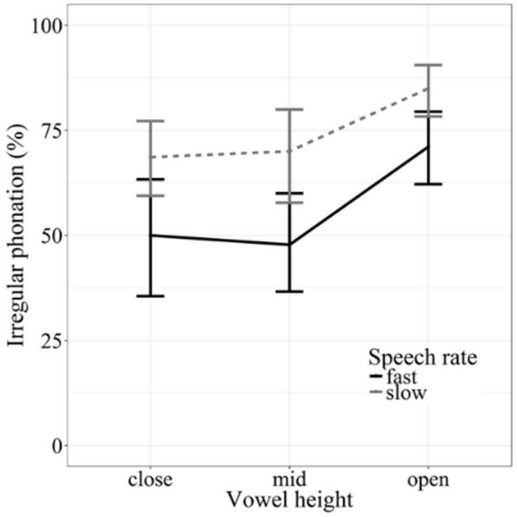


Figure 13.
The ratio of vowels produced with any kind of irregular phonation as a function of vowel height and speech rate (mean + 95% CI)

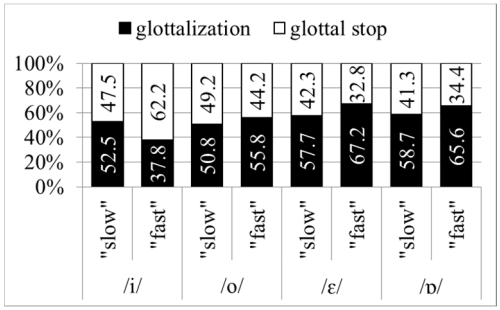


Figure 14.
The ratio of the two types of irregular phonation (relative to all irregular occurrences) as a function of vowel quality and speech rate

The ratios of glottalization and glottal stops were close to equal in the case of /i/ in the "slow" condition (52.5% vs. 47.5%, respectively), while in the "fast" condition the ratio of glottal stops (62.2%) was well above the ratio of glottalization (37.8%). For the "slow" condition the pattern was very similar in the case of /o/ (50.8% and 49.2% for glottalization and glottal stops, respectively); in the "fast" condition, however, glottalization (55.8%) was relatively more frequent than glottal stops (44.2%). In the case of /ɛ/ and /ɒ/, the ratio of glottalization exceeded the ratio of glottal stops in both of the conditions: it was 57.7% in /ɛ/ and 58.7% in /ɒ/ in the "slow" condition, and 58.7% in /ɛ/ and 65.6% in /ɒ/ in the "fast" condition. Although we observed differences in the glottalization to glottal stop ratio between the two conditions in three of the four analyzed vowels, the close /i/ showed the opposite tendency with the change in speech rate to that observed in the case of open vowels. The direction of the change in the mid /o/ was the same as in the case of the open vowels, but the degree of change was smaller.

3.3 Discussion

In Experiment 2 four vowel qualities of Hungarian embedded into real words were analysed in "slow" and "fast" speech rate conditions in terms of the frequency of occurrence of irregular phonation. We found significant effects of the *speech rate, vowel quality* and *vowel height* factors, while the factor *backness* (expressed in a twofold opposition) did not affect the data. The relative ratio of glottalization and glottal stops showed similar patterns as observed in Experiment 1. Thus we can conclude that in real Hungarian words embedded in non-facilitatory contexts of irregular phonation, the increased speech rate decreased the frequency of occurrence of vowels realized with irregular phonation in general, while the ratio of glottalization among these occurrences increased in all vowels but /i/ as speech rate increased.

4 General discussion and conclusions

In the two experiments reported in the present paper, we first tested the hypothesis that the frequency of occurrence of irregular phonation is higher in slow than in fast speech. Although, in this respect, the first experiment contradicted our expectations, in the second experiment, we eliminated some confounding factors (i.e., hiatus position) which may have affected the data in Experiment 1 in an undesired way, and we corroborated the hypothesis. Our second hypothesis may be regarded to be partially confirmed by the data: even though backness was not shown to have a significant effect on vowel-initial irregular phonation when it was expressed in the traditional twofold opposition (front vs. back) (in Experiment 1 and Experiment 2), the less traditional threefold analysis (front vs. central vs. back) revealed that there is an effect

observable in a numerically well-balanced comparison (in Experiment 1). According to this latter analysis, central and front vowels differed from back vowels in terms of their susceptibility to irregular phonation in slow speech. That is, while we observed a high number of vowel realizations with irregular phonation in basically all of the vowels in fast speech (irrespective of their backness), in slow speech, back vowels exhibited a much smaller number of realizations with irregularity in the voice source than central and front ones did. This effect, however, was observed only in Experiment 1, where the threefold recategorization was possible (as opposed to Experiment 2, where only the twofold-contrast categorization was attainable). In line with expectations, we also showed that open vowels favor irregular phonation more than mid and close ones both in slow and fast speech which finding corroborated our third hypothesis. Our fourth assumption claiming that faster speech rates reduce the relative amount of glottal stops, while increasing the frequency of occurrence of glottalization was not verified, since glottalization was more frequent in both speech rate conditions we studied. However, to some extent, at the different vowel heights studied, different tendencies were found. In Experiment 1, the vowel /u/, while in Experiment 2, the vowel /i/ appeared to be exceptions to some generally observed tendencies: in these vowels' cases the ratio of glottal stops was well above the ratio of glottalization, for /u/ in the "slow", while for /i/ in the "fast" condition. These results may suggest that the behaviour of close vowels is different from that of mid and open vowels with respect to the form of irregular phonation they elicit.

Considering that in the present study the effect of phonetic position, vowel quality, and speech rate were strictly controlled for and investigated, we can conclude that open vowels tend to elicit irregular phonation more than mid and close ones do, irrespective of backness. We can also conclude that the frequency of irregular phonation tends to be lower in fast than in slow speech (or at least in speech rate increased under laboratory conditions). The relative frequency of glottalization to glottal stops in phrase-initial position did not appear to be influenced by speech rate in general, which itself is inconsistent with the claims of earlier studies (e.g., Malisz et al., 2013). However, taking the analysed vowels separately into account, we observed that the behaviour of the close /i/ was opposite to that of the open /ɒ/ and /ɛ/. While the open vowels showed the widely documented tendency of being realized with relatively fewer glottal stops in fast speech, /i/ was produced with a relatively higher ratio of glottal stops under the same conditions. This result suggests that vowel height has an effect not only on the frequency of irregular phonation but also on the manner of its realization in the case of word- and phrase-initial vowels.

Acknowledgements

The authors are grateful to Gergely Varjasi for his valuable help in recruitment of the participants and in conducting the experiments.

References

Batliner, A., Burger, S., Johne, B., & Kiessling, A. (1993). MÜSLI: A classification scheme for laryngealizations. In *Working Papers, Prosody Workshop* (pp. 176-179). Sweden: Lund.

Bissiri, M. P., Lecumberri, M. L., Cooke, M., & Volín, J. (2011). The role of word-initial glottal stops in recognizing English words. In *Proceedings of Interspeech 2011* (pp. 165-168). Florence.

Boersma, P., & Weenink, D. (2016). *Praat: doing phonetics by computer.* Version 6.0.17. http://www.praat.org/

Bolla, K. (1995). *Magyar fonetikai atlasz. A szegmentális hangszerkezet elemei.* [Phonetic Atlas of Hungarian. The segmental entities.] Budapest: Nemzeti Tankönyvkiadó.

Bőhm, T., & Ujváry, I. (2008). Az irreguláris fonáció mint egyéni hangjellemző a magyar beszédben. [Irregular phonation as a speaker specific feature in Hungarian speech.] *Beszédkutatás 2008,* 108-120.

Dilley, L., Shattuck-Hufnagel, S., & Ostendorf, M. (1996). Glottalization of word-initial vowels as a function of prosodic structure. *Journal of Phonetics, 24,* 423-444.

Esling, J. H. (1978). The identification of features of voice quality in social groups. *Journal of the International Phonetic Association, 8,* 18-23.

Fletcher, J. (2010). The prosody of speech: Timing and rhythm. In W. J. Hardcastle, J. Laver, & F. E. Gibbon (Eds.), *The Handbook of Phonetic Sciences* (Second edition, pp. 521-602). Oxford: Blackwell.

Gósy, M. (2004). *Fonetika, a beszéd tudománya.* [Phonetics, the science of speech.] Budapest: Osiris Kiadó.

Henton, C., & Bladon, A. (1988). Creak as a sociophonetic marker. In L. Hyman, & Ch. N. Li (Eds.), *Language, speech, and mind* (pp. 3-29). London: Routledge.

Kassai, I. (1998). *Fonetika.* [Phonetics.] Budapest: Nemzeti Tankönyvkiadó.

Kohler, K. J. (2001). Plosive-related glottalization phenomena in read and spontaneous speech. A stød in German? In N. Grønnum, & J. Rischel (Eds.), *To Honour Eli Fischer-Jørgensen* (pp. 174-211). Kopenhagen: Reitzel.

Lancia, L., & Grawunder, S. (2014). Tongue-larynx interactions in the production of word initial laryngealization over different prosodic contexts: a repeated speech experiment. In S. Fuchs, M. Grice, A. Hermes, L. Lancia, & D. Mücke (Eds.), *Proceeding of the 10th ISSP* (pp. 245-248). Cologne.

Lefkowitz, D. (2007). *Creaky voice: Constructions of gender and authority in American English conversation.* Paper presented at American Anthropological Association. Washington, DC.

Mády, K. (2008). Magyar magánhangzók vizsgálata elektromágneses artikulográffal gyors és lassú beszédben. [Electromagnetic articulographic analysis of fast and slow Hungarian vowels.] *Beszédkutatás 2008,* 52-66.

Malisz, Z., Żygis, M., & Pompino-Marschall, B. (2013). Rhythmic structure effects on glottalisation: A study of different speech styles in Polish and German. *Laboratory Phonology, 4*(1), 119-158.

Markó, A. (2013). *Az irreguláris zönge funkciói a magyar beszédben.* [The functions of irregular phonation in Hungarian.] Budapest: ELTE Eötvös Kiadó.

Quené, H. (2007). On the just noticeable difference for tempo in speech. *Journal of Phonetics, 35*, 353-362.

Podesva, R. J. (2013). Gender and the social meaning of non-modal phonation types. In Ch. Cathcart, I-H. Chen, G. Finley, Sh. Kang, C. S. Sandy, & E. Stickles (Eds.), *Proceedings of the 37th Annual Meeting of the Berkeley Linguistics Society* (pp. 427-448.) https://journals.linguisticsociety.org/proceedings/index.php/BLS/article/view/832/615 (Retrieved: 30.01.2018)

Redi, L., & Shattuck-Hufnagel, S. (2001). Variation in the realization of glottalization in normal speakers. *Journal of Phonetics, 29,* 407-429.

Siptár, P., & Törkenczy, M. (2007). *The phonology of Hungarian.* Oxford University Press, New York.

Smith, A. (2010). Development of Neural Control of Orofacial Movements for Speech. In W. J. Hardcastle, J. Laver, & F. E. Gibbon (Eds.), *The Handbook of Phonetic Sciences* (Second edition, pp. 251-296). Oxford: Blackwell.

Stuart-Smith, J. (1999). Voice quality in Glaswegian. In J. J. Ohala (Ed.), *Proceedings of the XIVth International Congress of Phonetic Sciences* (pp. 2553-2556). San Francisco, 1-7 August 1999. Berkeley, Calif., USA: Linguistics Dept., University of California.

Szende, T. (1994). Illustrations of the IPA: Hungarian. *Journal of International Phonetic Association, 24*(2), 91-94.

Yuasa, I. P. (2010). Creaky voice: A new feminine voice quality for young urban-oriented upwardly mobile American women? *American Speech, 85*(3), 315-337.

WORD-INITIAL GLOTTAL MARKING IN HUNGARIAN AS A FUNCTION OF ARTICULATION RATE AND WORD CLASS

Tekla Etelka GRÁCZI[1,3] & Alexandra MARKÓ[2,3]
[1]Research Institute for Linguistics, Hungarian Academy of Sciences, Budapest, Hungary, [2]Department of Phonetics, Eötvös Loránd University, Budapest, Hungary, [3]MTA–ELTE "Lendület" Lingual Articulation Research Group, Budapest, Hungary
graczi.tekla.etelka@nytud.mta.hu, marko.alexandra@btk.elte.hu

Abstract
The present study's aim was to analyse the glottal marking of word-initial vowels in two speech styles, reading aloud and spontaneous speech, based on 12 speakers' Hungarian speech. In earlier studies of other languages (English, German and Polish), it was suggested that among several other factors, speech style, speech rate, vowel quality and word type had an effect on the glottal marking of word-initial vowels. In the Hungarian corpus analysed, speakers produced glottal marking significantly less frequently in spontaneous speech than in reading aloud. Both slower articulatory rate in the "carrier" pause-to-pause interval, and longer vowel duration went hand in hand with glottal marking. All the features of vowel quality were analysed, but only vowel height was found to play some minor role in glottal marking, while backness and rounding did not at all. In reading aloud, low and mid vowels were less frequently glottalized than high ones, while in spontaneous speech, high vowels were less frequently glottally marked than the other ones. The factor of vowel height played a significant role only in this latter speech style. Slower articulation proved to enhance the possibility of glottal marking in word-initial vowels in our study as well, in accordance with our hypothesis. With respect to word type (content vs. function words), our data imply that lexical prominence does not have an effect on the glottal marking of word-initial vowels in Hungarian. Alternatively, it can be supposed that phrase-initial position has a prevalent effect which overrides the effect of word type.

Keywords: glottal marking, word-initial vowels, reading aloud, spontaneous speech, articulation rate, word class

1 Introduction

Phonation results from vocal fold vibration. This vibration is usually quasiperiodic, leading to modal phonation; however, inconsistency may also appear. The phenomenon goes by several names, e.g., *irregular phonation, creaky voice, glottalization, laryngealization, vocal fry.* Whichever term is used, it is an umbrella term, as the vibration can be aperiodic in several ways: the timing or amplitude of adjacent periods may exceed the normal ranges of jitter and shimmer (e.g., Surana & Slifka, 2006), or their wide-spacing may be

DOI: http://doi.org/10.18135/CAPSS.75

atypical. For instance, Batliner et al. (1993) distinguished six types of laryngealization in approx. 30 minutes of spontaneous and read speech by four speakers. Dilley et al. (1996) studied texts read out by five speakers in radio news programs with respect to irregular voice quality (glottalization) occurring in word-initial vowels. They defined four types of realizations. Some researchers also analyse the glottal stop in this context (e.g., Dilley et al., 1996). Therefore it is not always clear which terms refer to which types of irregular phonation, or which terms are treated as synonyms.

The functions of irregular phonation vary across languages. In some languages, it expresses a phonological contrast – mostly it distinguishes pairs of sonorants from one another (for instance, in Mazateco, spoken in Mexico, it distinguishes vowels, and in some North-American Indian languages it distinguishes nasals); less frequently (e.g., in Hausa) distinguishing obstruents (see e.g., Gordon & Ladefoged, 2001). In several dialects of English, irregular phonation distinguishes allophones of syllable-final /t/ and /p/ (Pierrehumbert & Talkin, 1992).

Several researchers have investigated the role of irregular phonation in expressing emotions and/or attempted to use it in the automatic recognition of emotions (e.g., Batliner et al., 2007; Gobl & Ní Chasaide, 2003). The socio-cultural role of irregular phonation has also been demonstrated in an experiment involving young American women (Yuasa, 2010), and its conversational function has also been noted in English, where the realization of *yeah* with modal vs. irregular phonation is associated with distinct functions by the speaker (and the listener) (Grivičić & Nilep, 2004).

Studies revealed that prosodic factors are also linked to irregular phonation. Phrase boundary and/or stress elicit glottal marking at a higher rate (e.g., Umeda, 1978; Dilley et al., 1996; Rodgers, 1999; Kohler, 2001).

Rodgers (1999) found in German that glottalization is more frequent in content words than in function words (he defined content words as words with lexical meaning, and function words as words with grammatical meaning, based on Matthews, 1997) both in read and spontaneous speech. According to his interpretation, this difference may be partly due to the widely observed fact that content words can be accented in a relatively high ratio, while function words are typically unaccented. Rodgers (1999) also found that the frequency of occurrence of glottalization is higher in read speech than in spontaneous speech.

Speech rate also turned out to influence glottal marking: in slower speech, higher rates of glottal marking were observed than in faster speech (see e.g., Pompino-Marschall & Żygis, 2010; Malisz et al., 2013).

The occurrence of irregular phonation was found to show gender-related differences in many speech communities. A number of studies have found irregular phonation to predominate among male speakers (e.g., Esling, 1978 in

Edinburgh; Henton & Bladon, 1988 for speakers of RP and 'Modified Northern' English; Stuart-Smith, 1999 in Glasgow); however, the opposite tendency is also documented in the literature (e.g., Yuasa, 2010; Markó, 2013; Podesva, 2013).

It is also well known that the frequency of occurrence of irregular phonation is speaker dependent to a large extent. Some speakers hardly produce any irregular voicing, while some produce it fairly frequently (Umeda, 1978; Henton & Bladon, 1988; Dilley et al., 1996; Redi & Shattuck-Hufnagel, 2001; Slifka, 2006; for Hungarian: Bőhm & Ujváry, 2008; Markó, 2013). Therefore, the presence of glottalization has an eminent role in human speaker recognition (Bőhm & Shattuck-Hufnagel, 2007). It was also shown that the less speakers glottalize, the more probable it is that they do so at a phrase-, word- or vowel-to-vowel boundary position (Markó, 2013).

Kohler (2001, pp. 282-285) defined four types of glottalization covering "the glottal stop and any deviation from canonical modal voice" as follows.

(1) Vowel-related glottalization phenomena which signal the boundaries of words or morphemes beginning with vowels.
(2) Plosive-related glottalization phenomena [which] occur as reinforcement or even replacement of plosives.
(3) Syllable-related glottalization phenomena which characterize syllable types along a scale from a glottal stop to glottalization (e.g., Danish stød).
(4) Paralinguistic function of glottalization phenomena at the utterance level which comprise
    (i) phrase-final relaxation of phonation, and
    (ii) truncation glottalization, i.e., utterance-internal tensing of phonation at utterance breaks.

The present paper focuses on vowel-related glottalization phenomena (see above under (1)) in the case of word-initial vowels in Hungarian. The effect of vowel quality (especially vowel height) on the frequency of glottalization appears to prevail independently of the language, as it has a physiological background: in low/back vowels, the tongue is pulled back, and due to their mechanical links, the larynx is in a lower position (Moisik & Esling, 2011; Lancia & Grawunder, 2014). Therefore low/back vowels can elicit glottalization at a higher rate than close/front ones. This effect has also been demonstrated in Hungarian (Markó et al., 2018a).

The glottal marking of word-initial vowels was analysed in several earlier studies (e.g., Umeda, 1978; Rodgers, 1999; Redi & Shattuck-Hufnagel, 2001; Pompino-Marschall & Żygis, 2010) both in reading aloud and spontaneous speech, but mainly on English and German. More recently, Malisz et al. (2013) published a paper comparing German and Polish and highlighting the question if the rhythm characteristics of the given language have an effect on the appearance of glottal marking in the case of word-initial vowels. Both languages show vowel-related glottal marking, and neither of them treats the glottal stop as a

member of the phoneme inventory or an allophone. However, their rhythmical characteristics are different. Whereas German is unambiguously a stress-timed language with a mobile lexical stress pattern, the rhythmical status of Polish is debated (for details, see Malisz et al., 2013). Polish assigns fixed lexical stress to the penultimate syllable (with few exceptions), and as word stress is predictable, it is considered to be acoustically weak. According to the literature, in Polish, duration does not contribute to the expression of prominence, while in German word stress is represented primarily by duration at the lexical level. In the case of phrase level prominence, both languages apply changes in fundamental frequency, intensity and duration. While German is an inflected language, Polish is agglutinative with some inflecting characteristics. Malisz et al. (2013) summarized the relevant literature on glottal marking in German and Polish. In German, stressed and/or accented syllables, low vowels and content words (compared to function words) favour glottal marking. As for Polish, glottal stop was found in emphatic and boundary marking functions, the latter prevailing at the boundary of a prefix and a vowel-initial stem, and this pattern was more typical in the case of rare words than in frequent ones. In Polish, glottal marking was found to be more frequent in function words; however, it was not independent from prosodic phrase structure (with phrase-initial function words predominating in this pattern).

In the analysis of Malisz et al. (2013), both spontaneous and "prepared" speech were involved. The spontaneous subcorpus consisted of material from six instruction-givers of a Polish task-oriented dialogue corpus (202 word-initial vowels), and the utterances of four storytellers of a German spontaneous dialogue corpus (401 word-initial vowels). The gender of the speakers is not mentioned in the main text of the paper; however, Appendix A specifies that the German spontaneous subcorpus involved 2 female and 2 male speakers, compared to 3 female and 3 male speakers in the Polish spontaneous subcorpus. The "prepared" speech material consisted of public speeches of 4 Polish and 4 German "prominent speakers", mainly politicians, and all of them were males (472 Polish and 885 German word-initial vowels).

Polish speeches were found stress-timed, while Polish dialogues turned out to be syllable-timed. In the German material, both the prepared and the spontaneous subcorpora were measured as stress-timed. The results of the study showed that on word-initial vowels, glottal marking occurred more frequently in German (63.4%) than in Polish (45%). In German, glottal marking had a higher share in spontaneous speech (72.5%) than in prepared speech (59%), similarly to the situation in Polish (53.5% in spontaneous speech and 41.5% in prepared speech). In both languages, the majority of prominent vowels were glottally marked, and vowel height correlated with glottal marking, namely low vowels were marked glottally in a higher ratio. With respect to word type (content vs.

function words) opposite tendencies were found in the two languages. In Polish, function words were glottalized in a higher ratio (50% compared to 37.7% of content words), in contrast with German, where content words showed a higher ratio of glottal marking (77.3% compared to 57% of function words). However, in phrase-initial position, content words received more glottal marking than function words in the same position, even in Polish.

In the present paper, we analyse glottal marking in word-initial vowels in Hungarian spontaneous and read speech. Glottal marking is used here as an umbrella term in reference to any kind of irregularity in the vocal source, including glottal stop. By way of introduction, it is worth highlighting some relevant characteristics of Hungarian.

The Hungarian vowel system consists of 14 vowels which are paired in the dimension of quantity resulting in 7 short-long phonological pairs. However, the members of two short-long pairs (/ɛ/ and /eː/; /ɒ/ and /aː/) differ in their phonetic characteristics as well; therefore, from a phonetic point of view, 9 vowel qualities can be differentiated.

According to the traditional view, with respect to backness, /i y eː ø ɛ/ are considered as front vowels, while /u o ɒ aː/ are characterized as back vowels. It should be noted, however, that the status of the vowel /aː/ is ambiguous: while it is uniformly transcribed with the IPA symbol of a front vowel, it is generally classified as a back vowel (based on its morpho-phonological behaviour, namely its participation in vowel harmony) both by the phonological (e.g., Siptár & Törkenczy, 2000) and the phonetic literature (e.g., Kassai, 1998; Gósy, 2004). Note, however, that on the basis of an articulatory (X-ray) analysis of Hungarian vowels, Bolla (1995) claimed that /i eː ɛ/ are front vowels, /y ø aː/ are central vowels, while /u o ɒ/ are back vowels.

In the traditional view, again, Hungarian vowels are differentiated on the vowel height dimension as follows: /i y u/ are considered as close vowels, /eː ø o/ are categorized as close-mid, /ɛ/ is considered as open-mid, and /aː/ is considered as an open vowel (Kassai, 1998; Gósy, 2004). The short phonological counterpart of /aː/ in this view is considered to be the open-mid /ɔ/, while others (e.g., Mády, 2008) define the vowel at hand as an open /ɒ/. (In the present paper, we adhere to the latter notation and analysis.)

Hungarian is an agglutinative and syllable-timed language. At the lexical level, stress is highly predictable, assigned to the initial syllable of a content word, therefore word-level stress is non-distinctive (Szende, 1999). Lexical stress is considered to be expressed primarily by vowel duration (Szalontai et al., 2016; Mády et al., 2017). Hungarian is an obligatory syntactic focus marking language, which means (among other things) that in the case of narrow focus, the focused constituent shows the highest prominence, while the ensuing elements are deaccented. Narrow focus elements appear in particular syntactic

positions. Due to the close interrelations between syntax and accent distribution, several studies have argued that prosodic means do not play an important role in prominence marking in Hungarian, as suggested by evidence from both laboratory and spontaneous speech (Mády, 2012; Markó, 2012). However, some studies did find phonetic markers of focus prominence in Hungarian (Genzel et al., 2015; Szalontai et al., 2016; Mády et al., 2017). In particular, changes in fundamental frequency and duration were identified as strong predictors of prominence.

In Hungarian, various boundary marking functions of irregular phonation have been thoroughly investigated; however, many aspects of word-initial vowel-related glottal marking have not received a systematic analysis. The aim of the present study is to analyse some of the factors (speech style, vowel quality, speech rate, and word type) in Hungarian, which have been claimed to have an effect on the frequency of occurrence of vowel-related glottalization in word-initial position. Word type as a variable has never been introduced to the analysis of Hungarian before, and comparisons of spontaneous and read speech have not focused on vowel-related glottal marking. The effect of speech rate and vowel quality have only been analysed in systematically varied short stretches of laboratory speech (Markó et al., 2018a), with no pertinent data either from read speech or from spontaneous speech. Moreover, the results for German and Polish mentioned above (Malisz et al., 2013) invite cross-linguistic comparison.

Most of the studies of irregular phonation in Hungarian have used the spoken language database called BEA (see Gósy, 2012), which includes both spontaneous and read speech samples from the same speakers. The material of the present study also comes from this database (for details, see below).

Our hypotheses were formulated based on the previous literature on other languages, but not without taking into account the differences between Hungarian and languages that are well-studied in terms of glottal marking. With respect to speech style, in line with Rodgers (1999), our hypothesis was that the frequency of occurrence of glottal marking would be higher in read speech than in spontaneous speech. Read speech was expected to be more carefully produced, more fluent, and lacking hesitations. Besides, the text of read speech has a predetermined structure; therefore its organization might be more clearly marked by phonetic means such as glottal marking.

With respect to vowel quality, we analysed vowel height, backness and lip rounding separately. In line with earlier findings, we assumed that low vowels would show a higher rate of glottal marking than mid and high ones. Nevertheless, in terms of backness, previous results did not show an unambiguous pattern. In a study on Hungarian word-initial vowel-related glottalization (Markó et al., 2018a), the feature of backness was subjected to both twofold (front vs. back, see e.g., Gósy, 2004) and threefold comparisons

(front vs. central vs. back, see Bolla, 1995). While in the twofold comparison, the front vowels /i y eː ø ɛ/ and the back ones /u o ɒ aː/ were not distinguished by the frequency of glottal marking, the threefold comparison detected a significant difference between the front /i eː ɛ/ and the central /y ø aː/ versus the back /u o ɒ/ vowels, namely the last group elicited a higher ratio of glottal marking than the first two. Therefore, in the present study, /i y eː ø ɛ aː/ were considered as front vowels, while /u o ɒ/ were considered as back vowels (see also Gósy & Siptár, 2015).

Regarding the effect of speech rate, similarly to several earlier studies (e.g., Malisz et al., 2013), faster speech was assumed to reduce the frequency of glottal marking in word-initial vowels relative to slow speech.

In terms of word type, the previous results based mainly on German and Polish have mixed implications. Function words are non-prominent in general, and stress has been shown to correlate with glottal marking by several studies. Phrase-initial position, however, elicits glottalization at a higher rate. Considering that in Hungarian, the most frequent function words are the definite and indefinite articles (*a/az* 'the' and *egy* 'a(n)'), which begin with low/mid-low vowel and are typically positioned at the beginning of phrases (similarly to Polish), we did not expect word type to predict glottal marking.

2 Methods

2.1 Material

This study presents our results on read and spontaneous speech. The two subcorpora were chosen from the BEA database (Gósy et al., 2012). The database consists of various speech types, including both spontaneous and read speech. We analysed the spontaneous speech type, featuring texts in which the speakers talk about their lives, i.e. school years, jobs, hobbies, etc. This task involves a quasi-monologue, with the interviewer only asking questions if the subject seems to talk too little. As read material, we selected the sentence reading task of this database for our study. In this task, subjects read aloud 25 sentences of various lengths and syntactic structures; however, all are declarative sentences.

Texts added to the database are recorded under invariant circumstances. An AT4040 microphone is used, the speech is recorded digitally at 44 kHz and 16 bits.

All 25 sentences were labelled in the read speech material from each speaker, but in the case of very long spontaneous speech samples only the first appr. 4-5 minutes were labelled. This way, a total duration of 55.2 minutes of analyzed spontaneous speech was produced.

2.2 Subjects

The database collects speech of monolingual speakers of standard Hungarian. In Hungarian females' speech, glottalization was found more frequent than with male speakers of the same age groups (see Markó, 2013). Considering that gender may have an effect on glottal marking, the same number of female and male speakers (six from both genders) were chosen from the database: three young female (22 to 24 years), three young male (20 to 24 years), three middle age female (44 to 45 years) and three middle age male (39 to 45 years) speakers. None of them reported any hearing impairment or speech disorders. Both the reading and the spontaneous speech material of these speakers were used, i.e., the speakers in the two subcorpora were the same.

2.3 Data collection

Both the spontaneous speech samples and the readings were labelled in Praat (Boersma & Weenink, 2017). Three levels of labels were used (Figure 1). In the first tier, the pause-to-pause intervals of the subjects were transcribed for articulatory rate calculations. In the second tier, the words starting vowels of any kind were labelled. In these labels, the vowel quality, the word type (content or function word), and any possible further information were noted. In the third tier, each word-initial vowel was labelled and information on its glottal marking was included.

Cases where the glottal marking appeared only later, not at the start of the vowel, were not considered as being glottally marked due to their word-initial position. Besides, the cases where two phonemes of the same vowel quality met in a hiatus across word boundary and their boundaries could not be established were eliminated from the analysis. (For example: *ütötte el* /ytøtːɛ ɛl/ 'he spent (his time)'.

The following data were retrieved from the three labels by a Praat script:

(i) all words that start with any kind of vowel;
(ii) the type of the word (content or function word);
(iii) the initial vowel;
(iv) the duration of the initial vowel;
(v) the presence of glottal marking at the beginning of the vowel;
(vi) the duration of the relevant pause-to-pause intervals;
(vii) the quasi-phonetic transcript (one speech sound is represented by one character) of speech in the intervals.

In the analyses of vowel height, mid-low and low vowels were pooled (similarly to Markó et al., 2018a), therefore, we differentiated three levels of vowel height: high /i y u/, mid /eː ø o/ and low /ɛ ɒ aː/.

In the analysis of backness (based on the results of Markó et al., 2018a), /i y eː ø ɛ aː/ were considered as front vowels, while /u o ɒ/ were considered as back vowels in the present study. The vowel /aː/ is ambiguous regarding the feature of

backness, since /aː/ phonologically alternates with back /ɒ/; however, its phonetic position is central or front. As we hypothesized that glottal marking was related to articulatory phonetic properties, we analyzed /aː/ together with the front vowels (as we mentioned above).

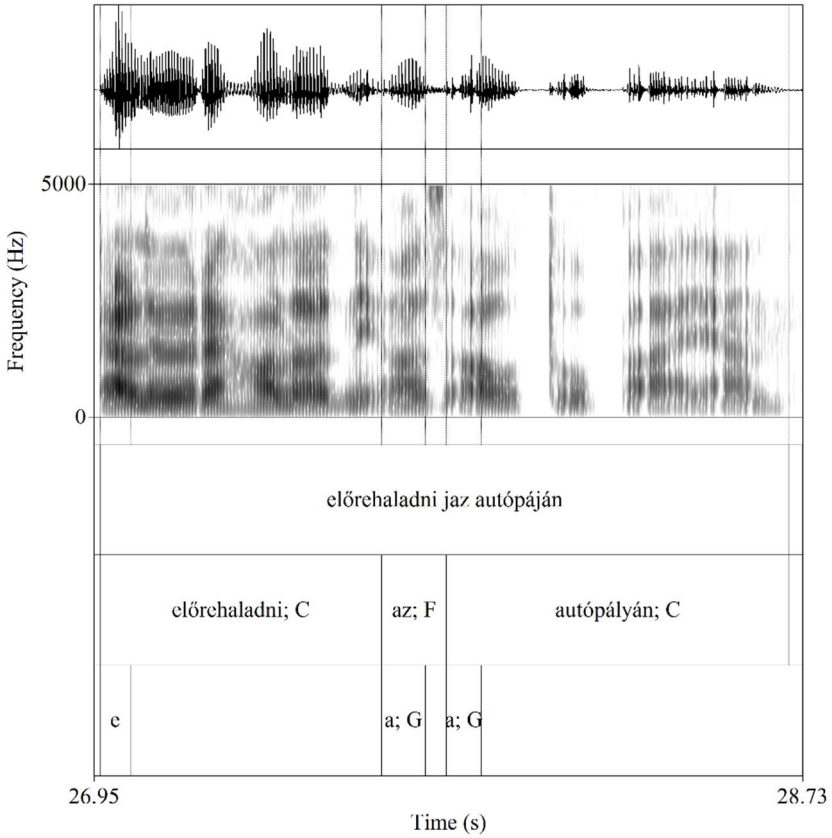


Figure 1.
Labeling sample. Reading task:
*A forgalom miatt csak nehezen lehetett **előrehaladni az autópályán.***
[ɛløːrɛhɒlɒdniʲ ɒz ɒutoːpaːjaːn]
'The traffic allowed only slow **proceeding on the highway.'**
(F = function word, C = content word; G = glottal marking)

The feature of lip rounding was analysed in accordance with the traditional view (e.g., Gósy, 2004): /y u ø o ɒ/ were considered as rounded, /i eː ɛ aː/ were considered as unrounded vowels. (Rounding partly co-varies with backness in the sense that all back vowels are rounded.)

Although Hungarian is a language with vowel quantity oppositions, we ignored this feature because of the skewed frequency of phonemes in this

language. In particular, high long vowels appear rarely in spontaneous speech and in non-phonetically compiled texts (e.g., Gósy, 2004). Also, quantity pairs of low vowels do not only differ in their duration but also in phonetic vowel height and other features. So, finally 9 vowel qualities were involved in the analyses, irrespective of their quantity (/i y eː ø ɛ aː u o ɒ/). However, the duration of the vowels was analysed with respect to the possible interrelation between duration and glottal marking.

2.4 Analyses and statistical methods

As noted in the Introduction, our question was whether and how speech style, vowel quality, word type and temporal factors affect the appearance of glottal marking in word-initial vowels. Therefore we analysed the interrelations of glottal marking with the quality and duration of the vowel, with word type, and with the articulatory rate of the carrier pause-to-pause interval in both speech styles. Only those intervals were used that included at least one word starting with a vowel. The analysis was carried out in two ways. As first we considered the measured articulatory rate as a scale variable, then we grouped the intervals in two ways: in a twofold and in a threefold comparison. In both comparisons only those pause-to-pause intervals were considered in which at least one word-initial vowel appeared (i.e., intervals in which all words started with consonants were excluded). In the twofold comparison, slow and fast categories were differentiated, while in the threefold comparison categories of (i) "slow", (ii) "medium" and (iii) "fast" were distinguished. In both the twofold and the threefold comparisons, the intervals belonging to different tempo categories were differentiated based on K-means Cluster analyses with iteration set to 20.

In order to test the statistical relevance of the factors in question, we carried out GLMM (General Linear Mixed Models), repeated measures ANOVA, Wilcoxon-test and Pearson's correlation. In order to test the effect of categorical factors on the appearance of glottal marking, the fixed factors in our GLMM model were *speech style, gender, vowel height, backness, rounding, word type,* and *articulatory rate clusters* (with both twofold and threefold clustering subjected to statistical analysis). In addition, these factors were also applied to the two speech styles separately. Scalar variables, i.e., *articulatory rate* and *vowel duration* were tested by GLMM, with glottal marking set as an independent variable. Repeated measures ANOVA was used to measure the statistical relevance of factors across and within speech styles.

Wilcoxon-test was used to compare the ratio of glottal marking in the three temporal clusters. Finally, Pearson's correlation was used to establish whether there was any correlation (i) between the average ratio of glottal marking and average articulatory rate of speakers, (ii) between average ratio of glottal marking and average vowel duration in the case of each speaker, and

(iii) between the ratios (in percentages) of glottal marking when the two speech styles are compared.

3 Results

3.1 General distributions in the corpora

Altogether 860 words starting with any kind of vowel were analysed in the read speech material of the 12 subjects, while 2288 items were found in their spontaneous speech samples. In the spontaneous subcorpus, the analysed vowels were glottally marked in 33.5±8.9%, while in the read speech material this ratio was higher: 54.1±11.5%. Although the distributions of the vowels and word classes analysed were different in the two speech styles, all speakers produced glottal marking more frequently in their read material than in spontaneous speech (the difference between the two speech styles varied between 11.4% and 60.6% speakerwise). This difference was significant ($F(1, 3130) = 44.594$, $p < 0.001$).

3.1.1 Interspeaker differences

Irregular phonation in general (not only in word-initial vowels) is known to be highly speaker- and gender-dependent. These patterns are also apparent in Hungarian (see e.g., Bőhm & Ujvári, 2008; Markó, 2013). Regarding the gender-specificity of the phenomenon, Markó (2013) found that female speakers produce this type of phonation more frequently than their male counterparts. The glottal marking of Hungarian word-initial vowels has not been analysed yet with regard to the subjects' gender. Therefore as a first step we checked the interspeaker and the gender-related variation of the data in our subcorpora.

The frequency of glottal marking in word-initial vowels varied between 26.9% and 74.0% in reading aloud, and between 17.5% and 54.5% in spontaneous speech. As already mentioned above, the standard deviations were not high and did not show large differences (reading: 11.5%, spontaneous speech: 8.9%), that is, interspeaker variability can be considered as moderate. Although the distribution of the various vowel qualities was not balanced within and across the subcorpora as a consequence of the inherent features of the two speech styles, speakerwise correlation was detected between the ratios of glottal marking in the two speech styles (Pearson's correlation: $r^2 = 0.606$, $p = 0.037$). This means that if a speaker produced less glottal marking in one speech style, s/he also produced less in the other one.

In order to see if there was any gender-specificity in our data, we calculated the mean and SD for the two gender groups (Figure 2). The female subjects produced glottal marking more frequently in both speech styles (reading aloud: 57.4±13.8%, spontaneous speech: 38.5±7.6%) than the male ones (reading aloud: 50.9±9.2%, spontaneous speech: 28.6±6.8%). In reading aloud, this

difference was somewhat lower than in spontaneous speech. The differences between the genders are significant in general ($F(1, 3130) = 44.594$, $p < 0.001$).

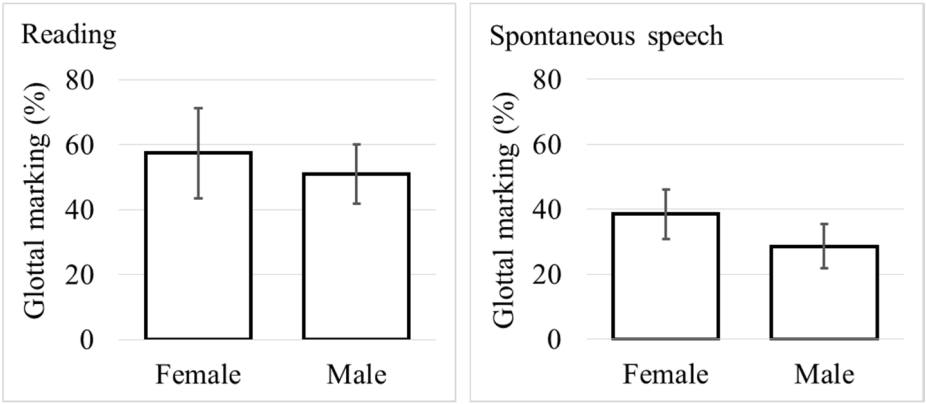


Figure 2.
Ratio of glottal marking in the word-initial vowels
as a function of gender
in reading aloud (left) and in spontaneous speech (right)

3.2 Vowel quality

3.2.1 Vowel height

Vowel height showed different tendencies in the two subcorpora (Figure 3). In reading aloud, the high vowels showed glottal marking in 60.2±16.1%, the mid ones in 61.7±18.6%, and the low ones in 52.5±11.8%. In spontaneous speech, the lowest ratio of glottal marking appeared in high vowels (26.2±7.1%), with mid and low ones not showing any difference (35.3±11.5%, 35.4±9.2%, respectively). Vowel height in itself did not have a significant effect on the frequency of glottal marking ($F(2, 3130) = 0.836$, $p = 0.434$) in general. However, this variation of the ratio of glottal marking varied significantly in the spontaneous speech subcorpus (repeated measures ANOVA: Wilks' $\lambda = 0.340$, $F(2, 11) = 9.713$, $p = 0.005$), while in reading aloud the results did not reveal significant differences among the analysed vowel categories (repeated measures ANOVA: Wilks' $\lambda = 0.594$, $F(2, 11) = 3.411$, $p = 0.074$). When analysing the results speaker by speaker, we can conclude that most subjects followed the tendency that was apparent in the given speech style, or showed no difference (< 10%) as a function of vowel height. In read speech, however, one person exceptionally departed from the general tendencies.

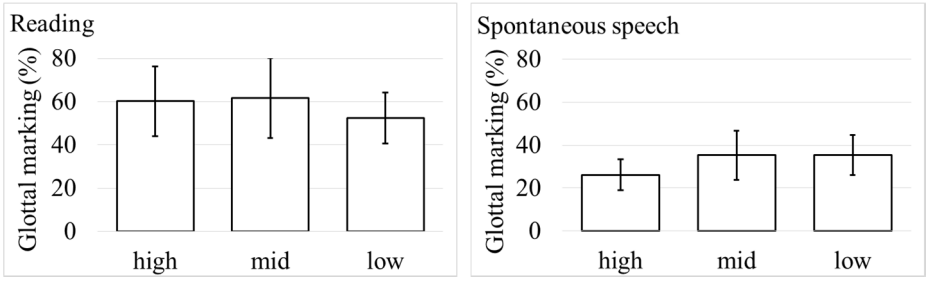


Figure 3.
Ratio of glottal marking in word-initial vowels
as a function of vowel height (mean and SD)
in reading aloud (left) and in spontaneous speech (right)

3.2.2 Vowel backness

The glottal marking of word-initial vowels did not show any tendencies related to vowel backness in either speech style. In reading aloud, three subjects produced glottal marking more frequently in back vowels, while five subjects produced more glottal marking in front vowels. In the case of four speakers, there was no difference (< 10%) in terms of frequency of glottal marking between the back and front vowels. In the spontaneous subcorpus, seven subjects produced more glottal marking in back vowels, three subjects in front vowels, and in the case of two speakers, the vowels did not show any difference (< 10%) in terms of glottal marking as a function of backness. There was no evidence of intraspeaker tendencies either. Speaker-specific differences led to a similar ratio of glottal marking in back and front vowels in both speech styles (Figure 4). In reading aloud, front vowels showed glottal marking in 55.3±7.3%), and back vowels in 53.3±15.5%. In spontaneous speech these ratios were 33.0±10.4% and 33.8±7.8%, respectively. These differences are not statistically significant (in general: $F(1, 3130) = 1.021$, $p = 0.312$, between the speech styles: $F(1, 3130) = 0.014$, $p = 0.907$).

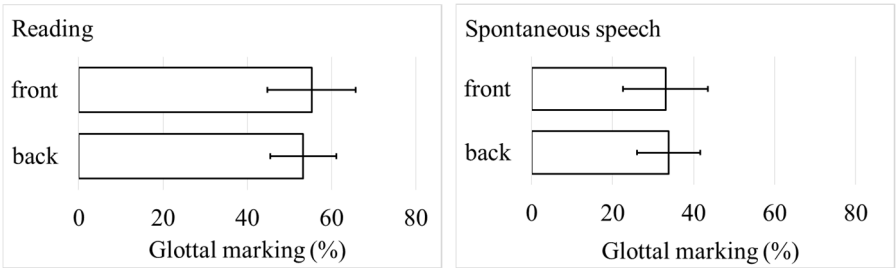


Figure 4.
Ratio of glottal marking in word-initial vowels
as a function of backness
in reading aloud (left) and in spontaneous speech (right)

3.2.3 Lip rounding

The effect of lip rounding on the glottal marking in word-initial vowels was also analysed. The mean data drawn from all speakers' results did not show any differences between the rounded and unrounded vowel groups in either of the speech styles (Figure 5.). In the read material, 55.3±7.1% of the unrounded vowels and 53.5±15.3% of the rounded ones showed glottal marking. In the spontaneous subcorpus these ratios were 33.2±10.3% and 33.6±7.7%, respectively. The intraspeaker differences were large in both speech styles. Some subjects used glottal marking more frequently in rounded, others in unrounded vowels, or the occurrence of glottal marking was similarly frequent in the two vowel classes. No tendency was found between the two speech styles: when a speaker glottally marked one kind of vowel (rounded vs. unrounded) more frequently in reading aloud, that did not mean that s/he glottally marked the same vowels more in his/her spontaneous speech as well. The results did not show any statistically significant difference (in general: $F(1, 3130) = 1.293$, $p = 0.256$, between the speech styles: $F(1, 3130) = 0.001$, $p = 0.971$).

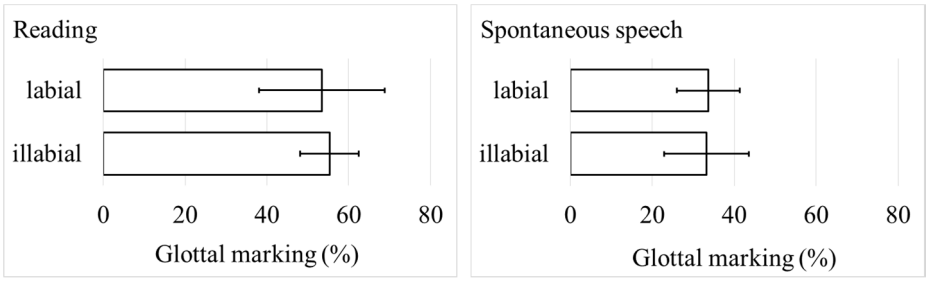


Figure 5.
Ratio of glottal marking in word-initial vowels
as a function of rounding
in reading aloud (left) and in spontaneous speech (right)

3.3 Temporal factors

3.3.1 Vowel duration

Our question about the relationship between vowel duration and glottal marking was whether the word-initial vowels that appeared with glottal marking were longer relative to the non-marked ones. As we hypothesized that slower articulation leads to higher frequency of glottal marking, here we did not separate the vowel qualities. The results are shown in Figure 6. In reading aloud, the glottally marked vowels' average duration was 83.9±10.6 ms, while the non-marked ones were shorter, 70.3±8.3 ms on average. In spontaneous speech, the mean duration of glottally marked vowels was 106.7±9.5 ms, while the non-marked ones were shorter again, with a mean of 80.7±7.0 ms. The general difference, namely glottally marked vowels being longer than non-marked ones,

was apparent in nine subjects' reading, and in the spontaneous speech of all the twelve participants. These differences were proven to be significant, suggesting that glottally marked vowels are longer than non-marked ones (in general: $F(1, 3144) = 54.524$, $p < 0.001$, between the speech styles: $F(2, 3144) = 6.404$, $p = 0.002$). Analysing the speech styles separately, we found that the difference in vowel duration was significant between the glottally marked and the non-marked vowels in both speech styles (reading aloud: Wilk's $\lambda = 0.415$, $F(2, 11) = 15.480$, $p = 0.002$; spontaneous speech: Wilk's $\lambda = 0.188$, $F(2, 11) = 47.363$, $p < 0.001$).

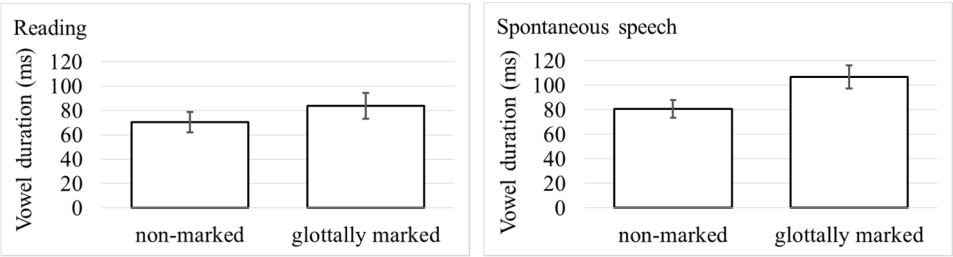


Figure 6.
Vowel duration as a function of glottal marking
in word-initial vowels
in reading aloud (left) and in spontaneous speech (right)

3.3.2 Articulatory rate

a) Absolute values of articulatory rate

Average articulatory rate. Our first question about the relationship between articulatory rate and glottal marking was if speakers with slower (average) articulatory rate produced more glottal marking in their word-initial vowels. The mean articulatory rate was 13.7 sounds/sec, and the standard deviation was 1.1 sounds/sec in both speech styles. Interspeaker variation was somewhat higher in spontaneous speech than in reading aloud, but still moderate. The subjects produced similar articulatory rates in the two speech styles; i.e., we were able to detect a strong correlation between the values in reading aloud and in spontaneous speech (Pearson's correlation: $r^2 = 0.801$, $p = 0.002$). As already described above in section 3.1, the appearance of glottal marking was 54.1±11.5% in reading aloud and 33.5±8.9% in spontaneous speech. As noted in section 3.1.1, this ratio also showed a certain correlation, i.e., the more frequently a speaker glottally marked their word-initial vowels in one speech style, the more they marked them in the other one as well (Pearson's correlation: $r^2 = 0.606$, $p = 0.037$). We hypothesized that a slower articulatory rate leads to a higher ratio of glottal marking in word-initial vowels. Based on the correlation data above, we might expect that the values of articulatory rate and frequency of

glottal marking measured subject-by-subject would also show correlation (Figure 7). In the read speech material, we found that the above mentioned tendency (that slower articulatory rate goes hand in hand with higher frequency of glottal marking) was present, however this correspondence still did not reach the level of significance ($r^2 = -0.574$, $p = 0.051$), which may imply that a larger corpus may prove this hypothesis. In any case, in our spontaneous subcorpus no significant relation was detected between the speakers' articulatory rate values and the appearance of glottal marking ($r^2 = -0.412$, $p = 0.183$).

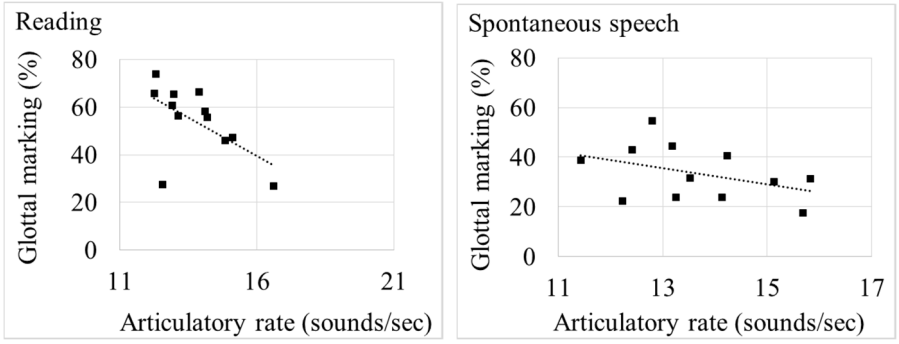


Figure 7.
Correlation of articulatory rate and ratio of glottal marking in word-initial vowels in reading aloud (left) and spontaneous speech (right)

Average articulatory rate as a function of voice quality. Our second question was whether the articulatory rate was slower in pause-to-pause intervals in which any glottal marking phenomenon appeared in the analysed vowels.

To answer this question, we compared the pause-to-pause intervals in which any of the word-initial vowels was pronounced with glottal marking and the ones in which there was at least one word starting with a vowel but none of them was glottally marked. In reading aloud, the mean of the articulatory rate was 13.9±1.0 sounds/sec in pause-to-pause intervals which did not show any word-initial vowel-related glottal marking, and 13.7±1.1 sounds/sec in those which contained at least one glottally marked word-initial vowel. In spontaneous speech, these results were 13.9±1.1 sounds/sec and 13.2±1.1 sounds/sec), respectively. Although the mean values apparently did not differ considerably, the statistical analysis revealed significant differences (in general: $F(1, 3144) = 121.231$, $p < 0.001$, between the speech styles: $F(2, 3144) = 48.240$, $p < 0.001$).

b) Slow vs. fast speech classification

After each pause-to pause interval with a word-initial vowel was detected and its articulatory rate was defined, the articulatory rate was categorized in two ways. The literature generally uses "slow" and "fast" speech categories in the presentation of results; therefore we also chose this classification as one method.

However, we also decided to perform a more detailed comparison, where "slow", "medium", and "fast" speech were considered. Thus we also applied a classification of the temporal data into three groups. The classifications were carried out by K-means cluster analysis in SPSS.

Twofold classification. The mean articulatory rate in the "fast" speech cluster was 15.2 sounds/sec in reading aloud and 15.0 sounds/sec in spontaneous speech, while "slow" tempo meant a mean of 12.5 sounds/sec in the first and 11.0 sounds/sec in the latter speech style. 54.0% of the word-initial vowels appeared in intervals considered as "slow" speech in reading aloud, while 36.9% of them in spontaneous speech. The glottal marking of word-initial vowels was more frequent in "slow" speech irrespective of speech style (Figure 8). In reading aloud, 59.4% of the word-initial vowels were glottally marked in "slow" speech, while 52.5% of them in "fast" speech. Interspeaker variability was higher in "fast" speech (SD = 18.3%, "slow" speech SD = 9.9%) in reading aloud. The difference between the two articulatory rates was larger with regard to the glottal marking of word-initial vowels in spontaneous speech (41.2% in "slow", and 30.9% in "fast" speech), and interspeaker variability was at 10.1% and 9.0%, respectively. The appearance of glottal marking was significantly different between the two temporal clusters when the two speech styles were considered together ($F(1, 3130) = 13.055$, $p < 0.002$). When reading aloud was analysed separately, "slow" and "fast" speech did not show any significant difference in terms of frequency of glottal marking (Wilk's $\lambda = 0.955$, $F(2, 11) = 0.517$, $p = 0.487$), while in spontaneous speech they did (Wilk's $\lambda = 0.467$, $F(2, 11) = 12.786$, $p = 0.004$).

Threefold classification. The mean articulatory rates in the three tempo clusters are presented in Table 1. The threefold temporal grouping revealed a significant effect of articulatory rate on the appearance of glottal marking in general ($F(2, 3130) = 5.627$, $p = 0.004$; between the speech styles: $F(2, 3130) = 6.308$, $p = 0.002$). In read speech, the frequency of word-initial vowel-related glottal marking did not vary with tempo changes (Figure 9), but was approximately 50% in all three clusters ("fast": 55.4%, "medium": 47.6%, "slow": 51.6%; Wilk's $\lambda = 0.882$, $F(2, 11) = 0.670$, $p = 0.533$). In addition, we observed that the slower the tempo, the higher interspeaker variability was (standard deviation: "fast": 9.2%, "medium": 10.2%, "slow": 20.0%). By contrast, spontaneous speech did show significant differences across the tempo clusters. In "medium" tempo, the ratio of glottally marked word-initial vowels was 24.2±9.3%, while in both the "fast" cluster and the "slow" one, a higher frequency of glottal marking was shown. The mean of the ratio of glottally marked word-initial vowels was 34.4±7.8% in "fast" tempo, and 50.8±18.1% in "slow" tempo, which was proven to be significantly different (Wilk's $\lambda = 0.300$, $F(2, 11) = 11.681$, $p = 0.002$). We tested the tempo clusters pairwise as well to see the differences in

detail. Wilcoxon-test was carried out and the *p*-value was expected to show significance below 0.01666 due to the threefold comparison. The "fast" cluster produced significantly less frequent glottal marking than the other two ones ("fast" vs. "slow": $Z = -2.667$, $p = 0.008$; "fast" vs. "medium": $Z = -2.824$, $p = 0.005$), while the "medium" and the "slow" tempi did not show any significant difference ($Z = -2.040$, $p = 0.041$).

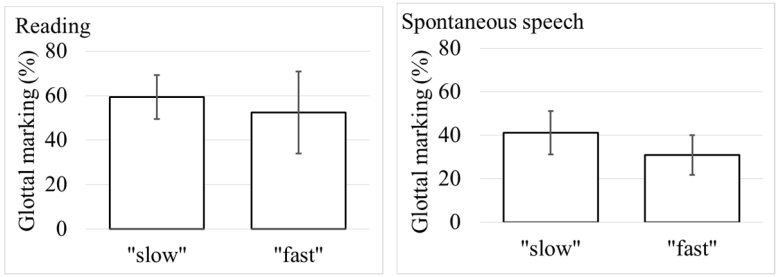


Figure 8.
Ratio of glottal marking in word-initial vowels as a function of articulatory rate clusters in reading aloud (left) and spontaneous speech (right)

Table 1. Mean articulatory rates in the three tempo clusters (sounds/sec)

	"slow"	**"medium"**	**"fast"**
reading aloud	10.6	13.3	16.6
spontaneous speech	9.0	13.1	15.6

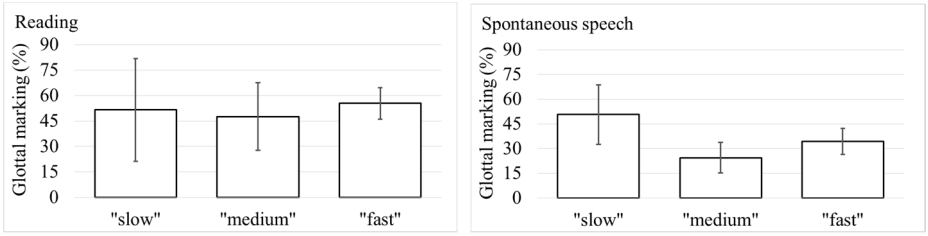


Figure 9.
Ratio of glottal marking in word-initial vowels as a function of temporal clusters in reading aloud (left) and spontaneous speech (right)

3.4 Word types

We compared the appearance of glottal marking in word-initial vowels between content and function words. We did not detect any differences between the two word types in either of the analysed speech styles. Some speakers tended to produce glottal marking more frequently in function words, some others in content words, and yet others produced glottal marking with the same frequency in both word types. In reading aloud, the frequency of glottal marking in the word-initial vowels of function words was 54.1±9.7%, and 54.4±12.3% in

content words; in spontaneous speech, the results were 33.8±8.9%, and 33.4±9.1%, respectively. The statistical analysis did not reveal any significant effect of word type on frequency of glottal marking ($F(1, 3130) = 0.401$, $p = 0.527$; between speech styles: $F(1, 3130) = 0.012$, $p = 0.912$).

4 Discussion and conclusions

The present study's aim was to analyse the glottal marking of word-initial vowels in two speech styles, reading aloud and spontaneous speech, based on 12 speakers' speech. In earlier studies of English, German and Polish, it was suggested that among several other factors, speech style, speech rate, vowel quality and word type had an effect on the glottal marking of word-initial vowels. In earlier studies, lower vowel height and slower articulation rate were found to enhance the likelihood of glottal marking, while speech style and word type had ambiguous effects (Rodgers, 1999; Malisz et al., 2013; Lancia & Grawunder, 2014). Our study raised the question if the phenomenon under study had similar facilitators in word-initial vowels in Hungarian.

In line with our first hypothesis (based on Rodgers' (1999) findings), speakers produced glottal marking significantly less frequently in spontaneous speech than in reading aloud. This difference might be traced back to the diverse speech planning strategies employed in reading aloud and spontaneous speech, which means among other things that prosodic marking is more direct in reading aloud due to a higher level of self-awareness in speech production.

The results of the German-Polish comparative study cited above (Malisz et al., 2013) showed that the glottal marking of word-initial vowels occurred more frequently in German than in Polish: 63.4% and 45%, respectively. In our material, we found glottal marking in 43.8% of word-initial vowels, which is close to the Polish data. In contrast with German and Polish, where glottal marking was more frequent in spontaneous speech (German: 72.5%, Polish: 53.5%) than in prepared speech (German: 59%, Polish: 41.5%), in Hungarian we found a higher ratio of glottal marking in read speech (54.1%) than in spontaneous speech (33.5%). Our results are in accordance with the assumption that read speech might be more clearly reflected by phonetic markers such as glottal marking.

We found (also similarly to earlier studies, e.g., Bőhm & Ujváry, 2008; Markó, 2013) that speakers producing more glottal marking in one of the speech styles also produced more glottal marking in the other style. This means that there was significant intraspeaker variation in the glottal marking of word-initial vowels between spontaneous speech and reading aloud, but the interspeaker differences did not correlate with speech style. Female speakers tend to produce more glottal marking in general – i.e., not only in word-initial vowels – in many

languages (e.g., in American English), also in Hungarian. This gender-specific difference was apparent in this specific phonetic position as well.

All the features of vowel quality were analysed, but only vowel height was found to play some minor role in glottal marking, backness and rounding did not at all. These latter effects were not analysed in the other languages in previous studies (to the authors' knowledge). In reading aloud, low and mid vowels were less frequently glottalized than high ones, while in spontaneous speech, high vowels were less frequently glottally marked than the other ones. The factor of vowel height played a significant role only in this latter speech style.

Slower articulation proved to enhance the possibility of glottal marking in word-initial vowels in our study as well, in accordance with our hypothesis. Both slower articulatory rate in the "carrier" pause-to-pause interval, and longer vowel duration went hand in hand with glottal marking. When dividing the articulatory rate into three clusters, we found that in the medium tempo word-initial glottal marking was less frequent. However, not only the "slow" but also the "fast" articulatory rate showed increased frequency of glottal marking in word-initial vowels.

In both German and Polish, the majority of prominent vowels are glottally marked. In the Hungarian corpus, utterance level prominence was not analysed due to its ambiguous phonetic characteristics. Even the literature on narrow focus, which is the most transparent phenomenon of prominence in Hungarian, has produced contradictory results (see Mády, 2012; Markó, 2012; Genzel et al., 2015; Szalontai et al., 2016; Mády et al., 2017), probably due to the syntactic determination of focus in Hungarian. The earlier studies which found phonetic correlates of focus prominence, found them in read speech, while the analysis of spontaneous speech did not reveal similar acoustic patterns. Moreover, in the read subcorpus of BEA, which was used for the present analysis, focus marking sentences were not included. Therefore, utterance level prominence could not be analysed unambiguously. Lexical level prominence, however, appears to be a more clearly definable phenomenon in Hungarian. Recently, several studies have proved that lexical prominence is expressed by durational differences (e.g., Szalontai et al., 2016; Mády et al., 2017; Markó et al., 2018b). At the lexical level, stress is highly predictable in Hungarian, assigned to the initial syllable of a content word. While content words bear first syllable stress, function words (e.g., definite and indefinite articles, postpositions, and conjunctions) appear as clitics, and do not bear stress. In contrast with many other languages, Hungarian is not assumed to display any covariance between word stress and vowel quality. Based on the distribution of lexical stress, we may assume that word types show interrelations with lexical prominence.

With respect to word type (content vs. function words), opposite tendencies were found in the languages mentioned above. In Polish, function words were

glottally marked in a higher ratio (50% compared to 37.7% of content words), whereas in German, content words showed a higher ratio of glottal marking (77.3% compared to 57% of function words). Pooled data (i.e., data not divided into read and spontaneous speech) for Hungarian showed the same ratios of glottal marking both in the case of content words (43.9%) and in function words (43.95%). As a function of speech type, we found that in reading aloud, the initial vowels of content words were also glottally marked at the same ratio as with function words: 54.4% and 54.1%, respectively. In spontaneous speech, the ratios were 33.4% for content words, and 33.8% for function words. These data imply that lexical prominence does not have an effect on the glottal marking of word-initial vowels in Hungarian. Or, as this factor was considered to raise an effect due to prosodic variation, there might be a difference with regard to this phenomenon in the languages. Another reason can also be assumed, similarly to Polish. Malisz et al. (2013) found a high number of phrase-initial function words in the Polish dialogue material, which is also the case in Hungarian (in both the read and the spontaneous subcorpora), due to the fact that the frequently used articles *(a, az, egy)* usually appeared in phrase-initial position. Therefore it can be supposed that phrase-initial position has a prevalent effect which overrides the effect of word type.

In our study, male and female speakers were involved in equal numbers, and the read material was the same in the case of all subjects. This contrasts with the German and Polish data, which represented different ratios of speakers in terms of gender (only males were involved in prepared speech, but both males and females in spontaneous speech), and the prepared material included various speeches. Although some of the effects have been found in all languages under analysis, the variability of the material and the relatively small number of the subjects might lead to ambiguous results. In several aspects (e.g., in the case of word type), language specific characteristics may be the reason behind the differences in the data. In order to support a better understanding of the universal and language specific elicitors of word-initial vowel-related glottal marking, future research will have to work with larger corpora which are also more amenable to crosslinguistic comparison.

Acknowledgements
The authors would like to thank Karolina Takács for participating in data labelling, Péter Siptár and the reviewers for their advising comments on the manuscript of the paper. The project was partially founded by the National Research Grant OTKA (project number: 108762).

References

Batliner, A., Burger, S., Johne, B., & Kiessling, A. (1993). MÜSLI: A classification scheme for laryngealizations. In *Working Papers, Prosody Workshop* (pp. 176-179). Sweden: Lund.

Batliner, A., Steidl, S., & Nöth, E. (2007). Laryngealizations and emotions: How many Babushkas? In M. Schröder, A. Batliner, & Ch. d'Alessandro (Eds.), *Proceedings of the International Workshop on Paralinguistic Speech* (pp. 17-22). (ParaLing'07, Saarbrücken 03.08.2007). Saarbrücken: DFKI. http://www5.informatik.uni-erlangen.de/Forschung/Publikationen/2007/Batliner07-LAE.pdf)

Boersma, P., & Weenink, D. (2017). *Praat: doing phonetics by computer* [Computer program]. Version 6.0. http://www.praat.org/

Bolla, K. (1995). *Magyar fonetikai atlasz – A szegmentális hangszerkezet elemei.* [Phonetic atlas of Hungarian – The segmental entities]. Budapest: Nemzeti Tankönyv-kiadó.

Bőhm, T., & Shattuck-Hufnagel, S. (2007). Listeners recognize speakers' habitual utterance final voice quality. In M. Schröder, A. Batliner, & Ch. d'Alessandro (Eds.), *Proceedings of the International Workshop on Paralinguistic Speech* (pp. 29-34). (ParaLing'07, Saarbrücken 03.08.2007). Saarbrücken. (http://www.bohm.hu/publications/BohmShattuckHufnagelParaling2007.pdf)

Bőhm, T., & Ujváry, I. (2008). Az irreguláris fonáció mint egyéni hangjellemző a magyar beszédben [Irregular phonation as an individual speaker's characteristic in Hungarian speech]. *Beszédkutatás 2008,* 108-120.

Dilley, L., Shattuck-Hufnagel, S., & Ostendorf, M. (1996). Glottalization of word-initial vowels as a function of prosodic structure. *Journal of Phonetics, 24,* 423-444.

Esling, J. (1978). The identification of features of voice quality in social groups. *Journal of the International Phonetic Association, 8,* 18-23.

Genzel, S., Ishihara, Sh., & Surányi, B. (2015). The prosodic expression of focus, contrast and givenness: A production study of Hungarian. *Lingua, 165,* Part B, 183-204.

Gobl, Ch., & Ní Chasaide, A. (2003). The role of voice quality in communicating emotion, mood and attitude. *Speech Communication, 40,* 189-212.

Gordon, M., & Ladefoged, P. (2001). Phonation types: a cross-linguistic overview. *Journal of Phonetics, 29,* 383-406.

Gósy, M. (2004). *Fonetika, a beszéd tudománya* [Phonetics: the science of speech]. Budapest: Osiris Kiadó.

Gósy, M. (2012). BEA: A multifunctional Hungarian spoken language database. *The Phonetician, 105-106,* 51-62.

Gósy, M., & Siptár, P. (2015). Abstractness or complexity?: The case of Hungarian /aː/. In É. Kiss, K., Surányi, B., & Dékány, É. (Eds.), *Approaches to Hungarian: Volume 14: Papers from the 2013 Piliscsaba Conference* (pp. 147-166). Amsterdam: John Benjamins Publishing Company.

Grivičić, T., & Nilep, Ch. (2004). When phonation matters: The use and function of yeah and creaky voice. *Colorado Research in Linguistics, 17*(1), 1-11. http://www.colorado.edu/ling/CRIL/Volume17_Issue1/paper_GRIVICIC_NILEP.pdf

Henton, C., & Bladon, A. (1988). Creak as a sociophonetic marker. In L. Hyman, C. N. Li (Eds.), *Language, speech, and mind* (pp. 3-29). London: Routledge.

Kassai, I. (1998). *Fonetika* [Phonetics]. Nemzeti Tankönyvkiadó, Budapest.

Kohler, K. J. (2001). Plosive-related glottalization phenomena in read and spontaneous speech. A stød in German? In N. Grønnum, & J. Rischel (Eds.), *To Honour Eli Fischer-Jørgensen* (pp. 174-211). Kopenhagen: Reitzel.

Lancia, L., & Grawunder, S. (2014). Tongue-larynx interactions in the production of word-initial laryngealization over different prosodic contexts: a repeated speech experiment. In S. Fuchs, M. Grice, A. Hermes, L. Lancia & D. Mücke (Eds.), *Proceedings of the 10th International Seminar on Speech Prosody (ISSP)* (pp. 245-248). Cologne.

Mády, K. (2008). Magyar magánhangzók vizsgálata elektromágneses artikulográffal gyors és lassú beszédben. *Beszédkutatás 2008,* 52-66.

Mády, K. (2012). A fókusz prozódiai jelölése felolvasásban és spontán beszédben [Prosodic marking of focus in read and spontaneous speech]. In M. Gósy (Ed.), *Beszéd, adatbázis, kutatások.* [Speech, data base, and research] (pp. 91-107). Akadémiai Kiadó, Budapest.

Mády, K., Reichel, U., & Szalontai, Á. (2017). A prozódiai prominencia (nem)jelölése a németben és a magyarban [The (non)-marking of prosodic prominence in German and Hungarian]. *Általános Nyelvészeti Tanulmányok, XXIX,* 77-98.

Malisz, Z., Żygis, M., & Pompino-Marschall B. (2013). Rhythmic structure effects on glottalisation: A study of different speech styles in Polish and German. *Laboratory Phonology* 4(1): 119-158.

Markó, A. (2012). A magyar hangsúly realizációinak és észlelésének összefüggése felolvasásban és spontán beszédben [Interrelations of realizations and perception of stress in Hungarian read and spontaneous speech]. In A. Markó (Ed.), *Beszédtudomány. Az anyanyelv-elsajátítástól a zöngekezdési időig* [Speech science: From first language acquisition to voice onset time] (pp. 277-303). ELTE BTK–MTA Nyelvtudományi Intézet, Budapest.

Markó, A. (2013). *Az irreguláris zönge funkciói a magyar beszédben.* [Functions of irregular voicing in Hungarian speech]. Budapest: ELTE Eötvös Kiadó.

Markó, A. Deme, A., Bartók, M., Gráczi, T. E., & Csapó, T. G. (2018a). *Speech rate (and vowel quality) effects on vowel-related word-initial irregular phonation in Hungarian.* In M. Gósy, & T. E. Gráczi (Eds.), *Challenges in analysis and processing of spontaneous speech* (pp. 49-73). Budapest: Research Institute for Linguistics of Hungarian Academy of Sciences.

Markó, A., Bartók, M., Gráczi, T. E., Deme, A., & Csapó, T. G. (2018b). Prominence Effects on Hungarian Vowels: A Pilot Study. In K. Klessa, J. Bachan, A. Wagner, M. Karpiński, & D. Śledziński (Eds.), *Proceedings of 9th International Conference on Speech Prosody 2018* (pp. 868-892). https://www.isca-speech.org/archive/SpeechProsody_2018/pdfs/138.pdf

Matthews, P. (1997). *The concise Oxford dictionary of linguistics.* Oxford: Oxford University Press.

Moisik, S., & Esling, J. H. (2011). The 'whole larynx' approach to laryngeal features. In *Proceedings of International Conference of Phonetic Sciences* (pp. 1406-1409). Hong Kong.

Pierrehumbert, J., & Talkin, D. (1992). Lenition of /h/ and glottal stop. In G. J. Doherty, & D. R. Ladd (Eds.), *Papers in laboratory phonology II: Gesture, segment, prosody* (pp. 90-117). Cambridge: Cambridge University Press.

Podesva, R. J. (2013). Gender and the social meaning of non-modal phonation types. In C. Cathcart, I-H. Chen, G. Finley, S. Kang, C. S. Sandy, & E. Stickles (Eds.), *Proceedings of the 37th Annual Meeting of the Berkeley Linguistics Society* (pp. 427-448).

Pompino-Marschall, B., & Żygis, M. (2010). Glottal marking of vowel-initial words in German. *ZAS Papers in Linguistics, 52,* 1-19.

Redi, L., & Shattuck-Hufnagel, S. (2001). Variation in the realization of glottalization in normal speakers. *Journal of Phonetics, 29,* 407-429.

Rodgers, J. (1999). Three influences on glottalization in read and spontaneous German speech. *Arbeitsberichte des Instituts für Phonetik und digitale Sprachverarbeitung der Universität Kiel, 25,* 173-280.

Siptár, P., & Törkenczy, M. (2000). *The phonology of Hungarian.* Oxford University Press, New York.

Slifka, J. (2006). Some physiological correlates to regular and irregular phonation at the end of an utterance. *Journal of Voice, 20*(2), 171-186.

Stuart-Smith, J. (1999). Voice quality in Glaswegian. In *Proceedings of the International Congress of Phonetic Sciences, 14,* 2553-2556.

Surana, K., & Slifka, J. (2006). Is irregular phonation a reliable cue towards the segmentation of continuous speech in American English? In *Proceedings of Speech Prosody 2006.* Dresden, Germany. http://20.210-193-52.unknown.qala.com.sg/archive/sp2006/papers/sp06_177.pdf.

Szalontai, Á., Wagner, P., Mády, K., & Windmann, A. (2016). Teasing apart lexical stress and sentence accent in Hungarian and German. In *Tagungsband 12. Tagung Phonetik und Phonologie im deutschsprachigen Raum* (P&P 12) (pp. 216-219).

Szende, T. (1999). Hungarian. In *Handbook of the International Phonetic Association. A guide to the use of the International Phonetic Alphabet* (pp. 104-107). Cambridge – New York – Melbourne – Madrid – Cape Town – Singapore – São Paolo: Cambridge University Press.

Umeda, N. (1978). Occurrence of glottal stops in fluent speech. *Journal of American Society of Acoustics, 64*(1), 88-94.

Yuasa, I. P. (2010). Creaky voice: A new feminine voice quality for young urban-oriented upwardly mobile American women? *American Speech, 85*(3), 315-337.

PHRASE-FINAL LENGTHENING OF PHONEMICALLY SHORT AND LONG VOWELS IN HUNGARIAN SPONTANEOUS SPEECH ACROSS AGES

Mária GÓSY & Valéria KREPSZ
Research Institute for Linguistics, Hungarian Academy of Sciences
gosy.maria@nytud.mta.hu, krepsz.valeria@nytud.mta.hu

Abstract
Phrase-final lengthening may concern the vowels or the consonants of the phrase-final syllable, or the whole syllable. Vowel quantity as a phonemic distinction was also shown to interact with phrase-final lengthening. In this study we sought to explore temporal patterns of phrase-initial, phrase-medial and phrase-final, phonemically short and long vowels ([ɔ], [aː]) in words with diverse numbers of syllables focusing on possible differences between Hungarian-speaking young and old subjects. Spontaneous narratives of 10 young and 10 old speakers were randomly selected from the BEA database. Both phonemically short and long vowels were significantly longer in phrase-final positions than in phrase-medial positions in both age groups. Durations of vowels in phrase-medial positions were significantly shorter than those occurring in phrase-initial positions in young speakers' speech while there were no differences in their durations between the two positions in old speakers' speech. All speakers preserved the durational differences in all positions to avoid violating the phonemic patterns of the vowel system. Word length had a significant effect on vowel durations.

Keywords: lengthening, spontaneous speech, [ɔ] and [aː] vowels, young and old speakers

1 Introduction

Phrase-final lengthening is a phenomenon that has been known in phonetics for several decades (e.g., Lindblom, 1968). For definition, the last syllable of the word is lengthened in a phrase-final position, at a prosodic boundary or before a phrase-final pause resulting in longer duration than that of a segmentally identical phrase-medial syllable. More than forty years ago Klatt, one of the first researchers of the topic, claimed that syllables or part of them were longer at the end of the sentences than those occurring in the middle of the sentences (1975). The phenomenon has been reported, on the basis of controlled experiments, to exist in various languages, irrespective of their typological and prosodic patterns, but also to show specific differences across languages (e.g., Beckman, 1992; Fletcher, 2010; Cho, 2016). Phrase-final lengthening has been confirmed

DOI: http://doi.org/10.18135/CAPSS.99

in numerous languages, for example in English (Turk & Shattuck-Hufnagel, 2007), Estonian (Krull, 1997; Plüschke & Harrington, 2013), Spanish (Oller, 1973; Rao, 2010), German (Kohler, 1983), as well as in Eskimo, Yoruba (Nagano-Madsen, 1992), Hebrew (Berkovits, 1993), Dutch (Cambier-Langeveld, 1997), Arab (de Jong & Zawaydeh, 1999), Chinese (Lee et al., 2004), Finnish (Nakai et al., 2009), Russian (Kachkovskaia, 2014), Japanese (Den, 2015), Chicasaw (Gordon & Munro, 2007); etc. Durational changes may concern the vowels or the consonant(s) of the phrase-final syllable, or the whole syllable (e.g., Berkovits, 1993; Turk, 2007; Dimitrova & Turk, 2012).

Diverse methods relating to subjects, speech materials and procedures can be found in the literature analyzing the phenomenon. The speech materials varied from diverse types of read speech (meaningless sound-sequences, words, sentences) to various types of spontaneous speech samples (narratives, conversations). Phrase-final lengthening has been studied in the speech of adults, children, bilinguals and in language learners' speech samples (both in their L1 and L2) as well as in the case of atypical speakers and even in motherese (Snow, 1994; Lieshout et al., 1995; Gerken, 1996; Baum, 1998; Byrd, 2000; Koponen & Lacerda, 2003; Hansson, 2003; Dankovičová et al., 2004; Byrd et al., 2006; Adam, 2014; Maastricht et al., 2016; etc.). Acoustic-phonetic analysis concern the final syllables, final segments, syllables and/or segments in various other phrasal positions (e.g., Cambier-Langeveld, 1997; Fougeron & Keating, 1997; White, 2002; Kachkovskaia, 2014), and apart from durational measurements, prosody has also been considered in various types of analysis (Wightman et al., 1992; Berkovits, 1994; Hofhuis et al., 1995; Cambier-Langeveld, 1997; Frota et al., 2007; Turk & Shattuck-Hufnagel, 2007; Frota, 2016).

Several factors are suggested that might trigger the lengthening of vowels like subglottal pressure, decreasing articulation activity, some kind of relaxation of articulation gestures, linguistic, phonological, and higher-level factors, as well as syntactic structures, syllable structures, local tempo modifications, speech melody effects, stress patterns, speech rhythm, etc. (e.g., Den, 2015). Some explanations focus on specific cognitive factors like speech planning strategies of the speaker, conscious boundary marking, or the disambiguation of the ambiguous sentences.

In Hungarian, a number of studies have dealt with the temporal properties of the phrase-final lengthening. The existence of the phenomenon was reported both in read (e.g., Kassai, 1982; White & Mády, 2008; Gósy & Krepsz, 2017a) and in spontaneous speech (Hockey & Fagyal, 1999; Markó & Kohári, 2015; Gósy, 2017; Krepsz, 2017). Results confirmed that phrase-final lengthening was more pronounced in spontaneous speech than in reading (Markó & Kohári, 2015). Since the vowel inventory of Hungarian contains vowel pairs distinguished by length (Siptár & Törkenczy, 2000), the question arose whether

speakers regulated their pronunciation in phrase-final syllables considering the phonemic quantity differences of vowels. Four pairs of vowels were selected in a study (Gósy & Krepsz, 2017b) that differed in phonemic length ([o], [oː], [i], [iː] vs. [ɔ], [aː], [ɛ], [eː]), the latter two pairs differed also in terms of vowel quality. Their durational differences were analyzed in phrase-initial, phrase-medial and phrase-final positions across increasing numbers of syllables of the words in read sentences Vowels were significantly longer in sentence final as opposed to medial positions, but no significant differences were found in the durations between initial and final positions. Although phonologically long vowels were significantly longer than phonologically short ones in all positions, sentence-final lengthening was more marked in the phonologically long than in the phonologically short vowels.

Two pairs of vowels differing phonemically ([o], [oː], [i], [iː]) were analyzed in spontaneous speech with the participation of young Hungarian-speaking subjects (Gósy, 2017). In addition, the durations of pauses following the phrase-final syllables were also analyzed. Results supported the previous findings with read sentences that vowels in phrase-final positions were significantly longer than those in phrase-medial positions. Vowels preserved the physical manifest-tations of their phonemic length differences in all phrasal positions, irrespective of vowel quality. Lengthening in phrase-final positions did not yield different ratios depending on phonemic length. There were, however, no interrelations in the durations of phrase-final vowels and the following pauses.

Finally, temporal patterns of syllables and consonants produced in phrase-final positions were analyzed compared to those occurring in phrase-initial and phrase-medial positions (Krepsz, 2017). Results showed that phrase-final length-ening exists not only in the case of vowels but also in that of consonants, and also for the whole syllable. The extent of the lengthening is heavily influenced by the quality of consonants and their position in the syllables. Since Hungarian is an agglutinating language with rich morphology, word length was also consid-ered in the analyses. The number of the syllables of the words was shown to have an effect on the segment and syllable durations in various phrasal positions.

In sum, research results concerning phrase-final lengthening and temporal patterns in Hungarian have confirmed that (i) the phenomenon exists both in read and spontaneous speech, (ii) phonemic length differences are preserved in physical durations in all phrasal positions, (iii) phrase-final lengthening concerns vowels, consonants and the whole syllable produced before pauses, and (iv) the number of the syllables of the words influences the extent of lengthening.

Age is acknowledged to be of considerable importance when speech is considered. Numerous studies have shown the effects of aging on various processes of speech production (Torre–Barlow, 2009). Age-related changes to speech are attributed to changes in the anatomy and physiology of the speech

mechanism, reduced auditory feedback, decreased accuracy of motor control, as well as modified psychic and cognitive functions (e.g., Liss et al., 1990; Wohlert & Smith, 1998; Degrell, 2000; Czigler, 2003; Xue & Hao, 2003; Burke & Shafto, 2004; Zraick et al., 2006; Torre–Barlow, 2009; Rodríguez-Aranda & Jakobsen, 2011).

The chronological age of 65 years is widely accepted as the beginning of the 'elderly' or 'older period of life' (see WHO proposal: www.who.int/healthinfo/survey/ageingdefnolder/en/). However, 'elderly', as an umbrella term, covers diverse periods and thresholds according to ages. In general, subjects of about 50 years of age are identified as middle-aged or pre-elderly while those falling between 60 and 74 years are called young-old. Subjects with ages between 75 and 90 years are the old-old people, and those older than 90 years are the oldest old or very old. The periods themselves show overlaps and shortcomings in referring naturally to a great variety of people of the same and similar ages. Aspects of chronological, biological, psychological, and social ages influence the age categories.

The age of a speaker can be predicted with fair accuracy by her/his speech properties including voice tremor, pitch, speaking rate, loudness, and fluency, etc. (Yorkston et al., 2010). With advancing age, speech changes in the accuracy of articulatory movements, fluency, and communicative effectiveness. The typical process of getting old has a natural effect on breathing, intensity of musculature used during speaking, articulation movements, the effectiveness of speech motor control, and planning of verbal utterances from several aspects (e.g., Enright et al., 1994; Berry et al., 1996; Bashore et al., 1998). A slowing of nerve conduction velocities in the peripheral nervous system and a decrease of central neurotransmitters is supposed to account for a general slowing of articulation in old speakers (Weismer & Liss, 1991). Although verbal communication experience of the elderly can play a role in speaking in old ages, there is also evidence that a general mechanism limits elderly speakers' speech performance.

Based on findings in the literature we can conclude that there are pronounced age differences in the timing of speech in pre-elderly and old speakers (e.g., Kail & Salthouse, 1994). Investigations confirmed that elderly speakers adjust the length and durational patterns of their utterances according to their physiological capacity (Winkworth et al., 1995). Old people's speech tempo was significantly slower than those of young speakers', they produced significantly shorter speech samples and slower articulation than young speakers did (Amerman & Parnell, 1992; Huber, 2008; Jacewicz et al., 2010; Bóna, 2012). Word durations of elderly people were reported to be significantly shorter than those of young speakers, while speech sound durations produced by old speakers were remarkably longer than those of young ones (Smith et al., 1987; Bóna, 2012, 2013; Kent, 2000; Fletcher & McAuliffe, 2015).

However, speech timing control of old speakers was shown to be the same as that of young speakers in temporal adjustments of consonant articulation according to consonant duration and vowel distance (Amerman & Parnell, 1992). Similar findings were reported on temporal patterns of VOTs produced by young and elderly people (Sweeting & Baken, 1982). Although no significant differences were found in VOT values between young and old speakers, variability of the data increased with age, both within subjects and between groups. Age did not appear to influence accuracy of temporal parameters in lip and jaw tracking (Ballard et al., 2001). Speech motor control exhibits inherent temporal properties of speech production, where some subprocesses and/or some local temporal organization may remain intact in aging, may be somewhat resistant to aging effects, or may employ specific strategies in old age (Brenk et al., 2009).

To our knowledge, few investigations were devoted to analysing phrase-final lengthening in the elderly. Swedish-speaking young (ages between 20 and 30 years) and old (ages between 55 and 75 years) speakers' spontaneous speech samples were examined according to the temporal patterns of phrase-final lengthening (Hansson, 2003). Findings showed no differences in the lengthening patterns depending on age. Speakers of the Chicasaw language were over 60 years old when their speech samples were examined, and they showed clear phrase-final lengthening according to their language specificity (Gordon & Munro, 2007). Studies on phrase-final lengthening in aphasic speech report data also of age-matched elderly controls where the latter show the phenomenon in contrast to aphasic patients (e.g., Hammond, 1990).

The question is whether the age of Hungarian-speaking adults is a decisive factor in phrase-final lengthening. How do old Hungarian speakers implement phonemic length differences and different word lengths when realizing phrase-final vowels as opposed to their realizations in initial and medial positions in spontaneous speech? Are there any differences in the elderly's temporal patterns when compared to those of young speakers? Do old speakers regulate utterance-final lengthening to preserve the phonemically relevant quantity of vowels? How do phrasal positions and word lengths influence the temporal patterns of vowels produced by old speakers? In this study, we seek to explore temporal patterns of phrase-initial, phrase-medial and phrase-final, phonemically short and long vowels ([ɔ], [aː]) in words with diverse numbers of syllables focusing on possible differences between young and old speakers. No one has analyzed phrase-final lengthening in Hungarian elderly speakers' speech so far.

Five hypotheses were formulated. (i) Phrase-final lengthening will preserve the phonemic quantity differences of the target vowels irrespective of the speakers' age, (ii) phrase-final lengthening would be less expressed in the old age than in the young, (iii) target vowels will not show durational differences in phrase-initial and phrase-medial positions in old speakers' speech, (iv) the

number of syllables of the words will influence the durations of the target vowels occurring in phrase-final positions, and (v) the length of words will have a greater effect on vowel durations as produced by old speakers than on those of young speakers.

2 Methodology

Spontaneous narratives (more than 9 hours' material) of 10 young subjects (aged between 20 and 30, mean: 25 years) and 10 old ones (aged between 70 and 80, mean: 75 years) were randomly selected from the BEA Spontaneous Speech Database of Hungarian (Gósy, 2012). The topic of the narratives was the same with all subjects, they spoke about their (past) jobs, families, everyday activities, hobbies. Each group consisted of an equal number of females and males. Speakers of the database spoke the Budapest dialect of Hungarian. Articulation and hearing were age-appropriate with all subjects, they did not encounter any articulation disorder or specific hearing loss.

A phonemic pair of short and long vowels ([ɔ], [aː]) was selected as target vowels (they are, however, different in tongue height and lip rounding). For this study, 3,672 vowels were segmented, of which 2,250 were phonologically short and 1,422 were phonologically long vowels. Young speakers produced 1,672 vowels while 2,000 vowels were produced by old people. The vowels occurred in phrase-initial (1,034 tokens), phrase-medial (1,412 tokens) and phrase-final positions (1,226 tokens) in the last syllables of the words (either in stems or in suffixes). Words varied according to their lengths, from disyllabic words to 6-syllable ones. Occurrences in stems and suffixes were very similar for both vowels and in all positions. Young speakers produced 283 short and 265 long vowels while old speakers produced 274 short and 211 long vowels. Both content words and function words were considered. Special attention was paid to the occurrences of the target vowels according to their phonemic length, the three phrasal positions, the number of syllables the words consisted of, and the ratios of stems and suffixes that contained them. All syllables containing the target vowels were unaccented irrespective of their phrasal positions.

The same target vowels as occurring in monosyllables were analyzed separately as a kind of control set. Altogether 1,033 such vowels were considered, of which 557 were phonologically short and 476 were phonologically long vowels.

Examples for [ɔ] (in orthography *a*) vowels both in stems and in suffixes (target vowels are written in bold, the abbreviation SIL stands for silence):

(1) SIL *magy**a**r szakra járok bár mostanában már úgy mesélem hogy alkalmazott nyelvészetre* SIL
'I study Hungarian language and literature though presently I say applied linguistics';
(2) SIL *ezzel a kiegészítéssel készen lesz a diplom**a*** SIL
'the thesis will be ready with this supplement';

(3) SIL *tudn**a**k tájékozódni három dimenzióban segítség nélkül* SIL
'they can orient themselves in all three dimensions without any help';
(4) SIL *akkor könyvkiadóban dolgozt**a**m és korrektori munkát végeztem* SIL
'I have worked in a publisher's office as a proof-reader'.

Examples for [a:] (in orthography *á*) vowels both in stems and in suffixes:

(5) SIL *a bar**á**t fontos minden gyereknek* SIL
'a friend is important for all children';
(6) SIL *sokan vannak akik dolgoznak de otthagyták mert ez a fajta ir**á**ny* SIL
'there are many people who are working [beside their studies] but they quit because this kind of way' SIL;
(7) SIL *a béték**á**n* [Bölcsészettudományi Karon] *magyar szakra járok* SIL
'I study Hungarian as my major at the ELTE university';
(8) SIL *arra gondoltam hogy ez a kirándul**á**s* SIL
'I have been thinking that this excursion'.

The speech material has been manually annotated by the two authors separately according to phrases (periods between two pauses), words and target vowels with simultaneous visual feedback in Praat software (Boersma & Weenink, 2015). The target vowels were segmented by defining the interval between the onset and offset of the second formants of the vowels. Segmentation was checked by a third phonetician (in case of disagreement, which was less than 1%, the vowels in question were excluded). A specific script was written for obtaining the values automatically. Examples for [a:] vowels in phrase-initial, phrase-medial and phrase-final positions produced by a male speaker are shown in Figure 1.

Durations of the vowels were analyzed according to (i) vowel quality, (ii) word length, (iii) phrasal position, and (iv) the speaker's age.

The distribution of the data were normal in both ages and in both vowels according to the Kolmogorov–Smirnov test (using SPSS 15.0 software). To test statistical significance, General Linear Mixed Models analyses were carried out to test the effects of the fixed factors *position, vowel quality, word length, age,* and their interactions on durations of the vowels (dependent factors). The confidence level was set at the conventional 95%.

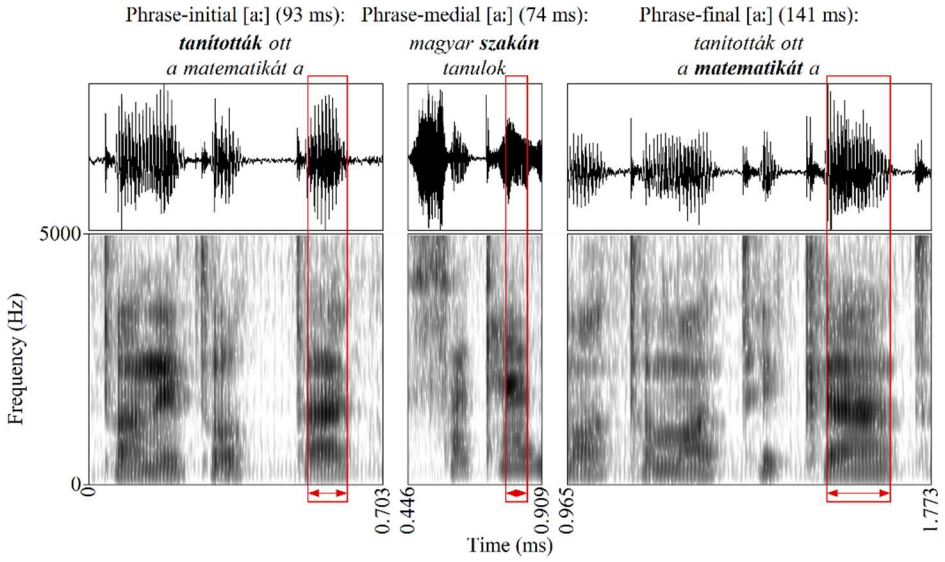


Figure 1.
Annotated samples of utterances containing the vowel [a:] in phrase-initial, phrase-medial and phrase-final positions (bold letters identify the carrier words of the target vowels)

3 Results

There will be two parts of the analysis in relation of the temporal patterns of the target vowels produced by both the young and old speakers. In the first part we will focus on the vowels that occurred in polysyllabic words while the other part is engaged with the vowels that occurred in monosyllables.

The presentation of the data will gradually and selectively be extended according to the factors that influence the durations of the target vowels as Figure 2 demonstrates. The data of the physical durations of the phonemically short and long vowels will be presented first. The next step concerns the distribution of the data according to the phrasal positions followed by extending the durational data of young and old adults, separately, on the one hand, and of the phonologically short and long vowels, separately, on the other. The next approach contains the data of the vowels' durations in the three phrase positions, distributed according to age and phonological length. The durational data according to word length are presented in relation to (i) phonological length, (ii) age, and (iii) interrelations of the two. Finally, the measured durations of the target vowels will be shown according to the phonological length of the vowels, the age of the speakers, the three phrasal positions and word length. Durational data of the target vowels occurring in monosyllables complete the presentation of the results.

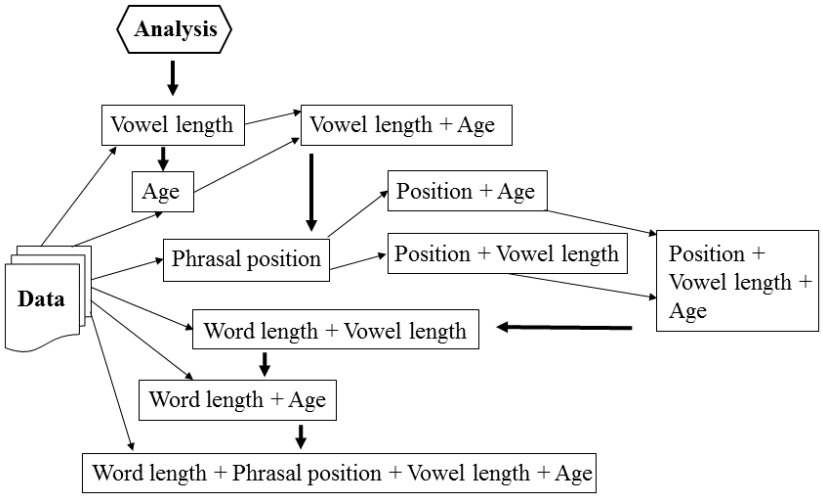


Figure 2.
Schematic process of the data analysis according to the factors involved

Different phonemic durations of [ɔ] and [aː] vowels are reflected by different durational values produced by both the young and old speakers (Figure 3).

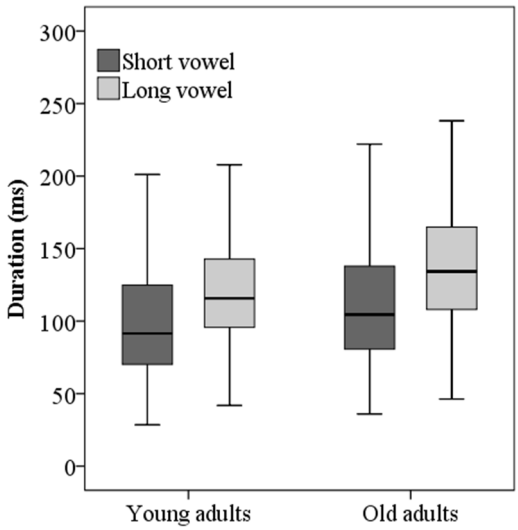


Figure 3.
Durations of the vowels analyzed as produced
by young and old speakers (medians and ranges)

Measured durations of [ɔ] vowels were significantly shorter than those of [aː] vowels. Mean durations of phonemically short vowels turned out to be 100 ms in

young and 113 ms in old speakers while those of phonemically long vowels were 123 ms in young and 139 ms in old speakers. The differences between the short and long vowels were significant in both age groups (for young speakers: $F(1, 1671) = 22.157$; $p = 0.011$; for old speakers: $F(1, 1999) = 17.844$; $p = 0.008$). In addition, the durational differences between the young and old speakers were also shown to be significant (for [ɔ] vowels: $F(1, 2249) = 22.157$; $p = 0.011$; for [aː] vowels: $F(1, 1421) = 22.157$; $p = 0.011$).

Vowel durations were analyzed according to the positions of the words in the utterance where they occurred in the last syllables of the words. Both phonemically short and long vowels were significantly longer in phrase-final positions than either in phrase-initial or phrase-medial positions irrespective of age groups (Figure 4). Values of all speakers confirmed the phenomenon of phrase-final lengthening in the cases of all vowels irrespective of their phonemic length. Durations of the vowels in phrase-medial positions (mean: 99 ms) were significantly shorter than those occurring in phrase-final positions (mean: 146 ms). Vowel durations were shorter in phrase-initial positions (123 ms) than in phrase-final but longer than in phrase-medial positions. Statistical analysis confirmed that position had a significant effect on the durations of the vowels ($F(2, 3670) = 51.389$; $p < 0.001$).

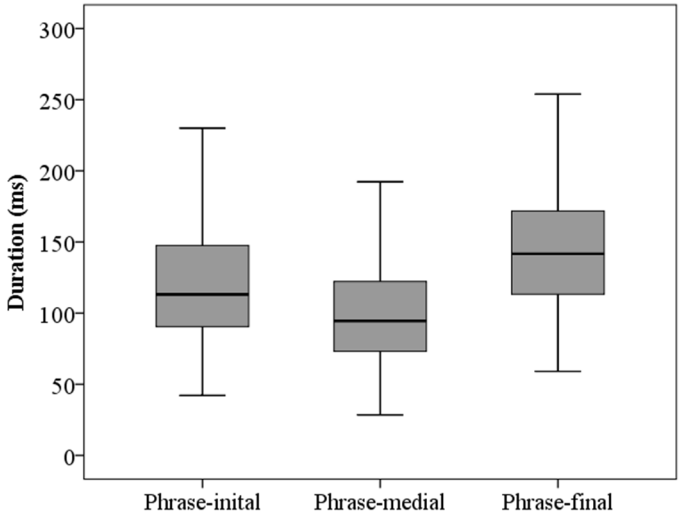


Figure 4.
Durations of the vowels analyzed in phrase-initial, phrase-medial and phrase-final positions (medians and ranges)

Durations of the vowels in the three phrase positions showed similar patterns produced by both young and old speakers (Figure 5). The mean duration of the vowels in phrase-initial position was 106 ms in young and 135 ms in old speakers, while the mean values were 100 ms and 98 ms, respectively, in phrase-medial positions. They turned out to be shorter than those occurring in phrase-final positions produced by young (mean: 142 ms) and by old speakers (147 ms). Statistical analysis confirmed that vowel durations were significantly different depending both on position and age (for young: $F(2, 1671) = 29.403$; $p = 0.014$; for old: $F(2, 1999) = 26.005$; $p = 0.018$).

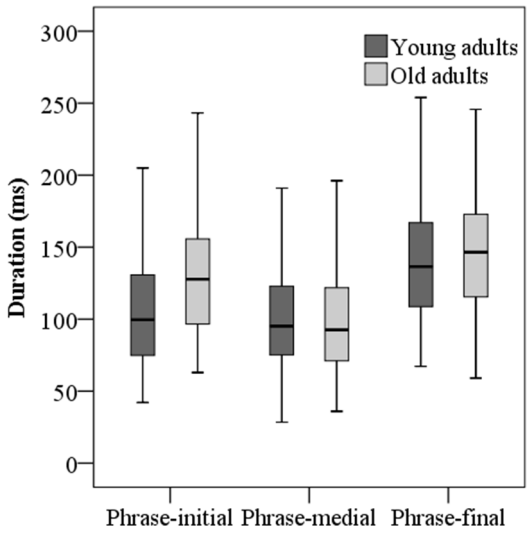


Figure 5.
Durations of the vowels analyzed in phrase-initial, phrase-medial and phrase-final positions depending on age (medians and ranges)

We analyzed the measured durations of the phonemically different vowels in the three positions but we did not consider the two age groups separately (Figure 6). As expected, the durations of the phonemically different vowels are similar across the three positions. Mean duration of [ɔ] vowels was 111 ms in phrase-initial positions, 89 ms in phrase-medial positions, and 138 ms in phrase-final positions. Mean duration of [aː] vowels was 139 ms in phrase-initial positions, 115 ms in phrase-medial positions, and 161 ms in phrase-final positions. The smallest difference in durations depending on the phonemic length of the vowels was found in phrase-final positions (23 ms, on average vs. 26 ms and 28 ms). Statistical analysis confirmed that the durations of both the phonemically short and long vowels are significantly different depending on position (for [ɔ]: ($F(2, 2249) = 11.678$ $p = 0.006$); for [aː]: ($F(2, 1421) = 13.408$; $p = 0.021$).

Data show that old speakers produced both phonemically short and long vowels longer than young speakers did. The mean value of [ɔ] vowels produced by young speakers was 100 ms in phrase-initial position, 89 ms in phrase-medial and 131 ms in phrase-final position. The mean value of the same vowels produced by old speakers was 113 ms in phrase-initial position, 88 ms in phrase-medial and 141 ms in phrase-final position. The temporal values of [aː] vowel produced by young speakers were 121 ms, 113 ms, and 157 ms, respectively while those produced by old speakers were 153 ms, 118 ms, and 165 ms, respectively. There was practically no difference found in the vowel durations between young and old speakers in phrase-medial positions. As expected from the former analyses, the mean durations of the phonemically long vowels indeed exceed all mean durations of the phonemically short ones in the same positions in both age groups (Figure 7). The lines in the figures demonstrate the age-specific differences in the target vowels' durations. Both the phonemically short and long vowels are longer in phrase-initial and phrase-final positions in old speakers than in young speakers, particularly in the intitial positions. However, the durational patterns are almost the same in phrase-medial positions.

Word length had a significant effect on vowel durations. Table 1 contains the durational values of the vowels analyzed, irrespective of position and age. The durations of [ɔ] vowels do not change dramatically according to the number of the syllables in the words. Their durations decrease by 18 and 20 ms between disyllabic words and words containing 5 and 6 syllables. The decrease of the durations of [aː] vowels according to the increasing number of syllables of the words is more remarkable, though the values vary. The temporal difference between phonemically short and long vowels decreases as word length increases (up to 29 ms as the largest difference).

The decrease of vowel durations according to the increasing length of the words can be experienced both with young and old speakers. What is interesting here is the different changes of the values in the two age groups. The durations are longer in old speakers' spontaneously produced words that contain 2, 3 or 4 syllables than the same ones produced by young speakers. However, old speakers' values abruptly shorten in words consisting of 5 and 6 syllables, and become shorter than those produced in the same word length by young speakers (Table 2).

Changes of vowel durations according to increasing word length seem to be in connection with syllable reduction in long words. Durations of the vowels depending on word length turned out to be signifiantly different in both age groups (young speakers: $F(5, 1671) = 30.214$, $p = 0.004$; old speakers: $F(5, 1999) = 19.789$, $p = 0.009$). The observed abrupt shortening in words consisting of 5 and 6 syllables in old speakers resulted in significant temporal differences between these long words and the shorter ones confirmed by post hoc tests (in

the case of 5-syllable words vs. shorter ones: $p = 0.040$, $p = 0.037$, $p = 0.031$, and 6-syllable words vs. shorter ones: $p = 0.032$, $p = 0.042$, $p = 0.021$).

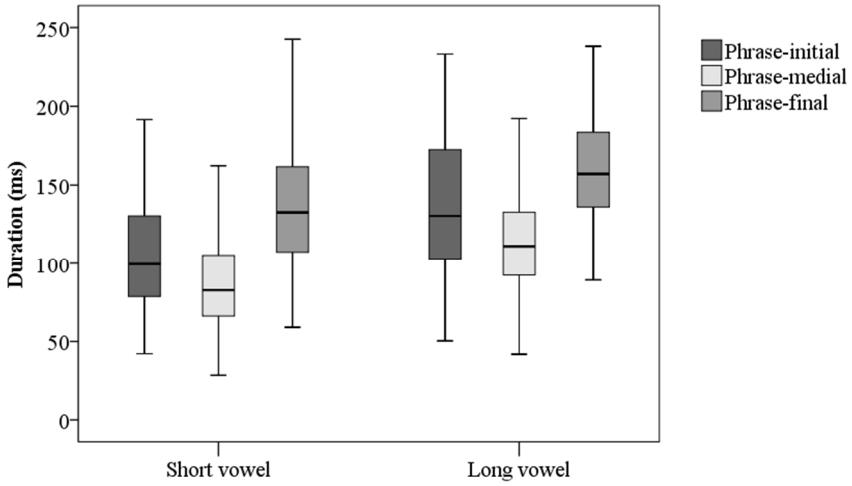


Figure 6.
Durations of [ɔ] and [aː] vowels in phrase-initial, phrase-medial and phrase-final positions irrespective of age (medians and ranges)

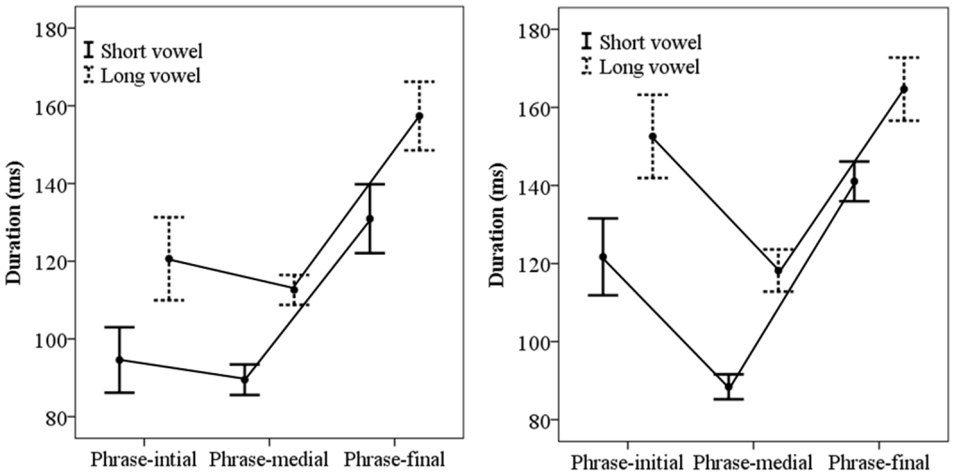


Figure 7.
Durations of the vowels analyzed in phrase-initial, phrase-medial and phrase-final positions produced by young (left) and old (right) speakers (means and ranges). Lines connecting the medians help to differentiate short and long vowels' durations

Table 1. Durations of the target vowels depending on the number of syllables in words (mean ± standard deviation)

Number of syllables in words	Durations of vowels (ms) [ɔ]	[aː]
2	111±40	129±40
3	109±45	133±41
4	103±55	122±39
5	91±38	131±41
6	93±31	104±19

Table 2. Mean durations of the vowels analyzed depending on words length and age of speakers (mean ± standard deviation)

Number of syllables in words	Durations of vowels (ms) young speakers	old speakers
2	112±40	125±41
3	110±42	123±46
4	104±38	116±55
5	107±51	97±35
6	102±32	89±20

Irrespective of phrase position, [ɔ] vowels are produced longer by old speakers than by young speakers with the only exception of the words consisting of 6 syllables. The vowels in these words were longer in the case of young speakers (Figure 8). The range of the vowel durations is wider than in the case of the young speakers, with the exception of the longest words.

Again, irrespective of phrase position, durations of [aː] vowels are similar to those of the phonemically short ones (Figure 9). Shortening of the durations can be seen in vowels produced by both young and old speakers; however, the decreasing tendency is more marked with the old speakers than with the young ones. There is no steep decrease in durational values for [aː] vowels in young speakers' speech. Statistical analysis confirmed that durations of both vowels depending on word length are statistically different in both age groups (see the summary in Table 3).

Finally, the analysis was extended considering all factors (Figures 10 and 11). Vowels were the longest in phrase-final positions, and reductions according to increasing word length are particularly characteristic in this position. The changes in the values are more marked with old speakers than with young speakers. Durations of [ɔ] vowels were the shortest in phrase-medial positions showing larger differences in old speakers' speech than in young speakers' speech. There were no statistically significant differences in durations between the vowels occurring in phrase-initial and phrase-medial positions in young speakers. The same differences in old speakers, however, proved to be significant ($F(5, 1998) = 23.589, p = 0.001$).

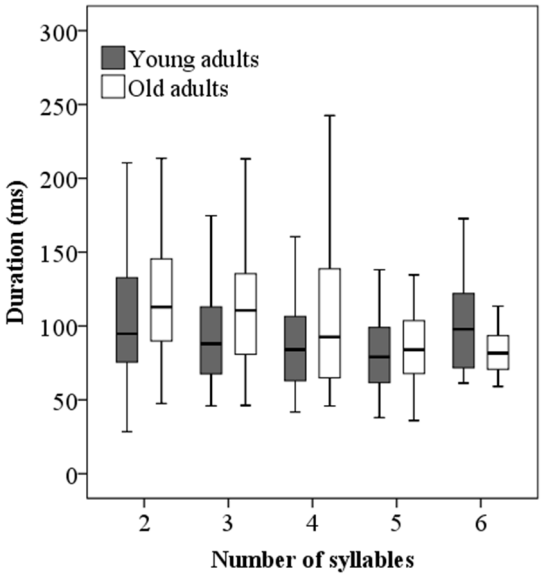


Figure 8.
Durations of [ɔ] vowels depending on word length produced by young and old speakers (medians and ranges)

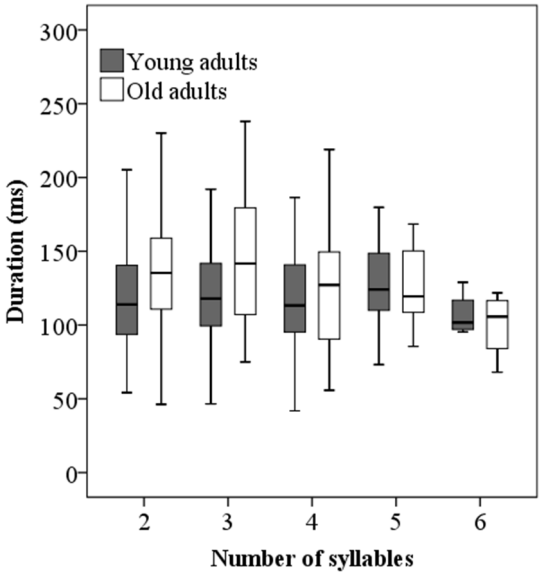


Figure 9.
Durations of [aː] vowels depending on word length produced by young and old speakers (medians and ranges)

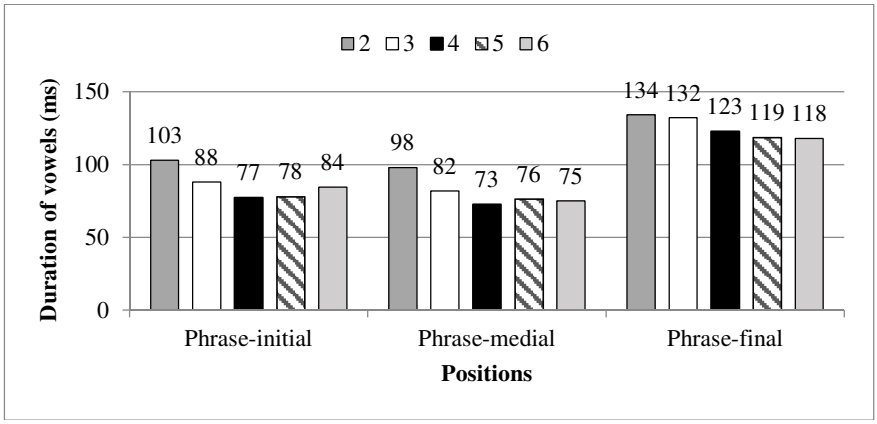


Figure 10.
Mean durations of [ɔ] vowels depending on word length in three positions produced by young speakers

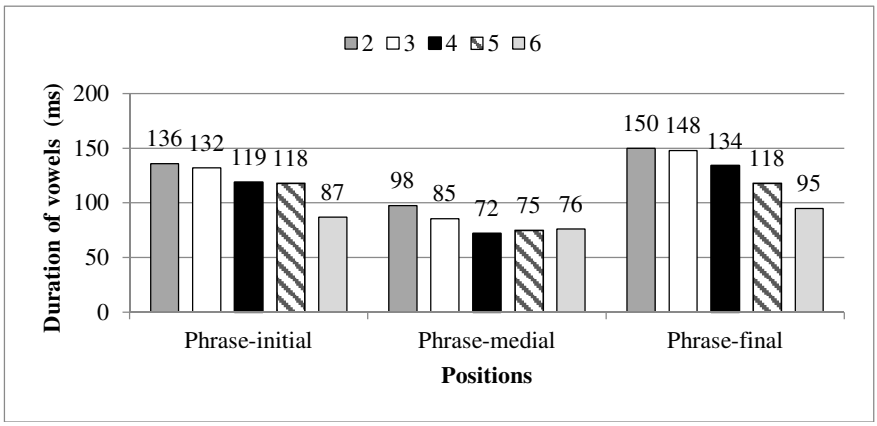


Figure 11.
Mean durations of [ɔ] vowels depending on word length in three positions produced by old speakers

The same analysis was carried out focusing on phonemically long vowels' durations considering words length, phrase position and age (Figure 12 for young and Figure 13 for old speakers). Values of [aː] vowels show similar distribution to what was experienced with phonemically short vowels. Vowels were longer in phrase-final positions than in phrase-medial positions in both age groups; however, durational patterns are different when considering all phrase positions. There were no statistically significant differences in durations of vowels occurring in phrase-initial and phrase-medial positions in young speakers' speech. On the contrary, it was between phrase-initial and phrase-final

positions that no significant differences were found in vowel durations in old speakers' speech.

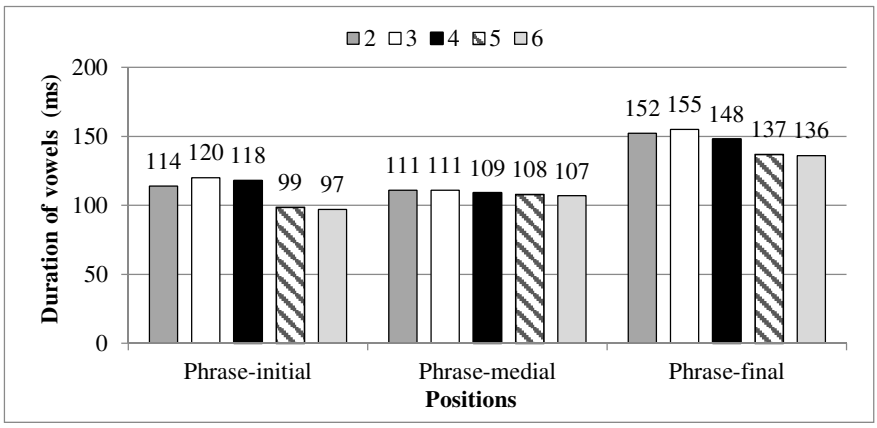


Figure 12.
Mean durations of [a:] vowels depending on
word length in three positons produced by young speakers

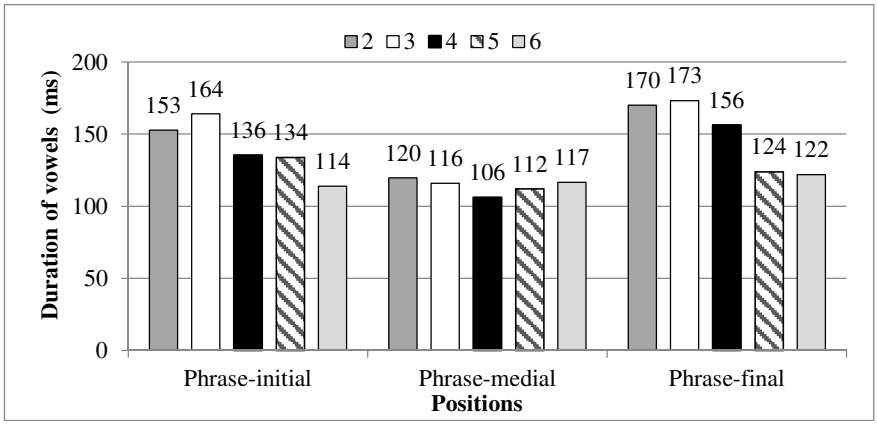


Figure 13.
Mean durations of [a:] vowels depending on
word length in three positions produced by old speakers

Results of the statistical analysis of the temporal patterns and their interactions for the vowels produced in polysyllabic words are summarized in Table 3. (The results of the detailed statistical analysis are given in the text.)

Table 3. Statistical data of the durational patterns of the vowels analyzed, as occurring in polysyllabic words (the value of df2 is 3.671 in all cases)

Factors	**df1**	***F*-value**	***p*-value**
position	2	51.389	< 0.001
vowel quality	1	88.932	< 0.001
word length	4	7.336	< 0.001
age	1	10.934	0.001
positon * word length	8	2.481	0.006
position * age	2	12.967	< 0.001
vowel quality * word length	4	7.762	< 0.001
position * vowel quality * age	2	6.676	0.001
vowel quality * word length * age	4	3.426	0.004
position * vowel quality * age * word length	6	3.341	0.002

3.1 Temporal patterns of monosyllables

Monosyllables are characteristically longer than the syllables of polysyllabic words (e.g., White & Mády, 2008; Gósy & Krepsz, 2017a). Therefore, it is expected that vowels of monosyllables should also be longer than those occurring in longer words. In addition, all monosyllabic content words have word stress (at least theoretically), while the syllables we measured in this study so far have definitely no perceivable lexical stress (according to the authors' judgement). Considering all these facts, we decided to pay special attention to vowel durations occurring in monosyllables in both age groups, and analyzed them separately from those in polysyllabic words. Table 4 summarizes the descriptive data.

All phonemically long vowels produced by both the young and old speakers were significantly longer than the phomemically short vowels. All vowels produced by old speakers were significantly longer than those produced by young ones. Phonemically short vowels produced both by young and old speakers show quasi-regular changes according to phrase positions: those occurring in phrase-medial positions were the shortest while those occurring in phrase-final positions were the longest. Durations of vowels occurring in phrase-initial positions fall in between (Figures 14 and 15).

Statistical analysis of the target vowels occurring in monosyllables confirmed significant differences in durations of vowels (for young speakers: $F(1, 547) = 27.155$; $p = 0.016$; for old speakers: $F(1, 484) = 19.306$, $p = 0.011$). In addition, phrase position also proved to have a significant effect on durations (young speakers, [ɔ] vowels: $F(1, 282) = 17.033$, $p = 0.018$ and [aː] vowels: $F(1, 264) = 19.345$, $p = 0.017$; old speakers, [ɔ] vowels: $F(1, 273) = 10.414$, $p = 0.021$ and [aː] vowels: $F(1, 210) = 20.686$, $p = 0.012$).

Table 4. Mean durations and standard deviations (mean ± SD) of the target vowels in monosyllables depending on position and age

Position	Mean duration of vowels (ms)			
	Young adults		Old adults	
	[ɔ]	[aː]	[ɔ]	[aː]
phrase-initial	95±64	117±26	148±81	176±45
phrase-medial	89±50	96±29	139±78	151±56
phrase-final	136±67	151±33	181±59	205±51

Temporal patterns of both the phonemically short and long vowels in monosyllabic and polysyllabic words are similar in young adults. The mean values of the short vowels are longer in monosyllables than those occurring in polysyllabic words; however, no differences were found between them occurring in phrase-medial positions. Durations of phonemically long vowels are longer in polysyllabic words in all positions although the differences are not large in all cases. The temporal differences of the target vowels between phrase-medial and phrase-final positions are more marked in monosyllables than in polysyllabic words in young adults.

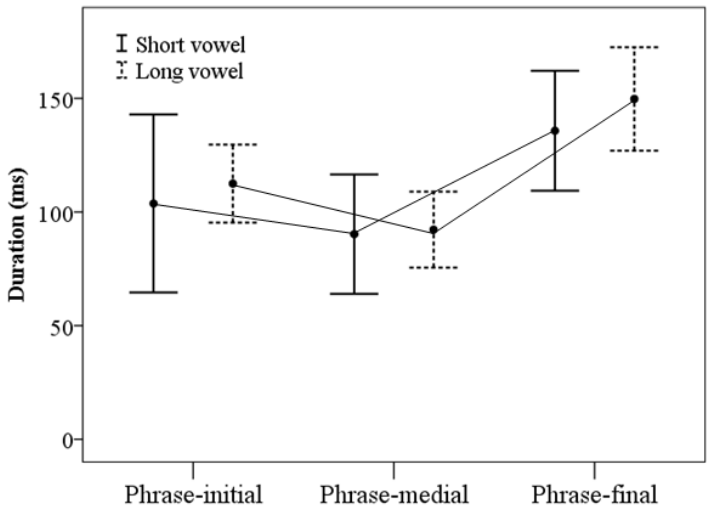


Figure 14.
Durations of the target vowels that occur in monosyllables depending on phrase position in young speakers (means and ranges)

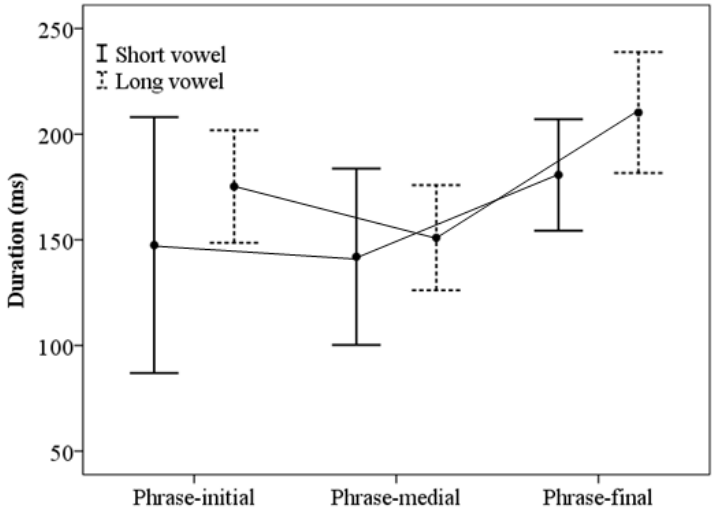


Figure 15.
Durations of the target vowels that occur in monosyllables depending on phrase position in old speakers (means and ranges)

The durational values of the target vowels produced by old speakers are longer in monosyllables than in polysyllabic words without exception. Phrase-final lengthening is more marked in phonemically long vowels in monosyllables while it is more marked in phonemically short vowels in polysyllabic words. The strength of phrase-final lengthening (expressed in longer durations compared to phrase-medial position) seems to be different depending on age.

4 Conclusions

Several questions were raised concerning phrase-final lengthening in spontaneous speech, in view of the agglutinating nature of Hungarian. We wanted to obtain evidence for (i) phrase-final lengthening using two vowels differing in phonemic quantity ([ɔ], [aː]), (ii) preservation of the target vowels' phonemic quantity in phrase-final positions, (iii) the durational differences of the target vowels depending on phrasal positions, (iv) the effect of word length on the target vowels' durations, and (v) the assumed differences in temporal patterns betwen young and old speakers differing by 50 years, on average.

Our results confirmed again – in accordance with the former research results (e.g., Gósy & Krepsz, 2017a) – that utterance-final lengthening does exist in Hungarian spontaneous speech. As expected, both phonemically short and long vowels were significantly longer in phrase-final position than in initial and medial positions irrespective of age. This means that old speakers' production exhibits the same effect on vowel durations in the phrase-final positions as was

observed in young speakers. We can conclude that phrase-final lengthening is a phenomenon that is characteristic of speech (and language) but not of adult age. So, our first hypothesis was confirmed.

Our findings supported the claim that phonemic vowel quantity contrasts are preserved in all phrasal positions, including phrase-final ones. The clear distinction between short and long vowels also in phrase-final positions suggests that speakers avoid violating the phonemic patterns of the vowel system. Speakers preserved the phonological quantity differences of the target vowels in all phrasal positions meaning that the measured durations of the phonologically long vowels were longer than those of the phonologically short ones. So, our hypothesis was also confirmed here. We hypothesized that phrase-final lengthening would be less expressed in the old age group than in the young one. Findings did not support this assumption: Temporal patterns of phrase-final lengthening showed similar tendencies in the cases of both the young and old speakers.

Results showed that significantly different durations were produced by the young and the old speakers. The target vowels of the polysyllabic words were longer in old than in young speakers irrespective of the phonemic length differences of the vowels. This can obviously be explained by the old age: the relatively slow articulation gestures and slow cognitive operations of the old speakers. It has often been noted that older adults used slower speaking rates (e.g., Shipp et al., 1992; Winkler et al., 2003; Bóna, 2013).

Vowels were the shortest in phrase-medial positions and longest in phrase-final positions in both the young and old age groups. However, differences were found in the measured durations between the phrase-initial and phrase-medial positions depending on age. The durational differences of the target vowels in these two positions were less large as produced by young speakers than in those produced by old speakers. We suggest that old speakers seem to signal the phrase-initial position to a larger extent than did the young speakers. The reason behind this temporal difference may be in connection with the old speakers' supposed intention to mark the beginning of their phrases. However, further research can confirm or reject this assumption. We hypothesized that target vowels would not show durational differences in phrase-initial and phrase-medial positions in old speakers' speech. This assumption, however, was not confirmed.

The temporal patterns in relation to the number of the syllables of the words showed similar tendencies in old speakers as it was found with the young participants. The slight differences concern the reduction patterns. The reductions of the vowel durations according to the increasing length of the words showed both increased and decreased mean values in the case of young speakers, particularly in phrase-initial and phrase-medial positions. The reductions are more gradual according to the increasing number of syllables in the words in the case of old speakers, particularly in phrase-final positions. Old speakers scarcely

reduce the vowel durations in the other two phrasal positions. We think that the reduction differences depending on age are in connection with slower articulation and slower high-level operation of speech planning with old speakers. The similar durations of the target vowels in phrase-initial and phrase-medial positions as well as the decrease of durations along with the increase of word length in old speakers' speech is assumed to be the consequences of both their breathing and cognitive processing (Hooper & Cralidis, 2009).

We hypothesized that old speakers would reduce their vowel durations in the phrase-final positions in long words more than young speakers would do. The data supported this assumption in the case of words consisting of 5 and 6 syllables. We suggest that physiological constraints of the old speakers would result in the need of reduced articulation of final vowels of the long words. The question is, however, whether accessing lemmas or the whole phonological forms prior to articulation takes longer time for old speakers that requires in some sort of fast finishing the articulation of the long words. Or, it is just the necessary articulation of 5 and 6 syllables without breaking off as it is possible between two shorter words in connected speech (breathing capacity, control over the structure of the long words, specific articulation strategy of elderly speakers, cf., Brenk et al., 2009).

The target vowels in monosyllables were significantly longer than in polysyllabic words produced by old speakers. This finding can be explained by the different lexical access of short and long words, on the one hand, and by some time gaining behavior of the old speakers they use in the cases of the monosyllables. We suggest that the different temporal patterns of the target vowels depending on phrasal positions between young and old speakers is the result of the old speakers' intention to mark largely the phrasal positions. Old speakers want to be understood more than young speakers do.

We conclude that clear distinction of short and long vowels also in phrase-final positions suggests that speakers avoid violating the phonemic patterns of the vowels irrespective of age. Speech motor control refers to the systems and strategies that control the production of speech (Kent, 2000), and this control works throughout the speaker's lifespan. The temporal patterns analyzed in this research show some age-specific differences along with the preservation of the phonological representations of the target vowels and their physical realizations.

The findings of our research raise further questions on the possible effect of individual speech planning and articulation on the temporal patterns of the phrase-final syllables of words across ages. This requires further investigations.

Acknowledgments
We wish to thank Péter Siptár as well as the anonymous reviewers for their help with an earlier version of this paper. The research was supported by the National OTKA 108762 Project.

References

Adam, H. (2014). Dysprosody in aphasia: An acoustic analysis evidence from Palestinian Arabic. *Journal of Language and Linguistic Studies, 10,* 153-162.

Amerman, J. D., & Parnell, M. M. (1992). Speech timing strategies in elderly adults. *Journal of Phonetics, 20,* 65-76.

Ballard, K. J., Robin, D. A., Woodworth, G., & Zimba, L. D. (2001). Age-related changes in motor control during articulator visuomotor tracking. *Journal of Speech, Language and Hearing Research, 44,* 763-777.

Bashore, T. R., Ridderinkhof, K. R., & van der Molen, M. W. (1998). The decline of cognitive processing speech in old age. *Current Directions in Psychological Sciences, 6,* 163-169.

Baum, S. R. (1998). The effects of utterance length on temporal control in aphasia. In *Proceedings of International Congress of Acoustics (ICA).* Seattle, 1998. http://www.icacommission.org/Proceedings/ICA (Retrieved 14.02.2014.)

Beckman, M. E. (1992). Evidence for speech rhythms across languages. In Y. Tohkura, E. Vatikiotis-Bateson, & Y. Sagisaka (Eds.), *Speech perception, production and linguistic structure* (pp. 457-463). Oxford: IOS Press.

Berkovits, R. (1993). Utterance-final lengthening and the duration of final-stop closures. *Journal of Phonetics, 21,* 479-489.

Berkovits, R. (1994). Durational effects in final lengthening, gapping, and contrastive stress. *Language and Speech, 37,* 237-250.

Berry, J. K., Vitalo, C. A., Larson, J. L., Patel, M., & Kim, M. J. (1996). Respiratory muscle strength in older adults. *Nursing Research, 45,* 154-159.

Boersma, P., & Weenink, D. (2015). *Praat: doing phonetics by computer.* www.praat.org

Bóna J. (2012). A rövid-hosszú magánhangzók realizációi idősek spontán beszédében. [Realizations of short vs. long vowels in old speakers' spontaneous speech]. *Beszédkutatás 2012*, 1-15.

Bóna J. (2013). *A spontán beszéd sajátosságai az időskorban* [*Characteristics of spontaneous speech in old age*]. Budapest: ELTE–Eötvös Kiadó.

Brenk van, F., Terband, H., Lieshout van, P., Lowit, A., & Maassen, B. (2009). An analysis of speech rate strategies in aging. In *Proceedings of Interspeech 2009* (pp. 792-795).

Burke, D. M., & Shafto, M. A. (2004). Aging and language production. *Current Directions in Psychological Sciences, 13,* 21-24.

Byrd, D. (2000). Articulatory vowel lengthening and coordination at phrasal junctures. *Phonetica, 57,* 3-16.

Byrd, D., Krivokapić, J., & Lee, S. (2006). How far, how long: on the temporal scope of prosodic boundary effects. *Journal of the Acoustical Society of America, 120*(3), 1589-1599.

Cambier-Langeveld, T. (1997). The domain of final lengthening in the production of Dutch. In J. Coerts, & H. de Hoop (Eds.), *Linguistics in the Netherlands* (pp. 13-24). Amsterdam: John Benjamins.

Cho, T. (2016). Prosodic boundary strengthening in the phonetics-prosody interface. *Language and Linguistics Compass, 10,* 120-141.

Czigler, I. (2003). Időskori kognitív változások: Pszichofiziológiai megközelítés [Cognitive changes in old age: Psychophisiological approach]. In Cs. Pléh, Gy. Kovács, & B. Gulyás (Eds.), *Kognitív idegtudomány* [*Cognitive Neuroscience*] (pp. 343-355). Budapest: Osiris Kiadó.

Dankovičová, J., Pigott, K., Wells, B., & Peppé, S. (2004). Temporal markers of prosodic boundaries in children's speech production. *Journal of the International Phonetic Associaion, 34,* 17-36.

Degrell, I. (2000). A központi idegrendszer változásai öregedésben [Changes in central nercous system in aging]. In I. Czigler (Ed.), *Túl a fiatalságon. Megismerési folyamatok időskorban* [*Beyond youth. Cognitive processes in old age*] (pp. 11-130). Budapest: Akadémiai Kiadó.

Den, Y. (2015). Some phonological, syntactic, and cognitive factors behind phrase-final lengthening in spontaneous Japanese: A corpus-based study. *Laboratory Phonology, 6,* 337-379.

Dimitrova, S., & Turk, A. (2012). Patterns of accentual lengthening in English four-syllable words. *Journal of Phonetics, 40,* 403-418.

Enright, P. L., Kronmal, R. A., Manolio, T. A., Schenker, M. B., & Hyatt, R. E. (1994). Respiratory muscle strength in the elderly. *American Journal of Respiratory and Critical Care Medicine, 149,* 430-438.

Fletcher, J. (2010). The prosody of speech: Timing and rhythm. In W. J. Hardcastle, J. Laver, & F. E. Gibbon (Eds), *The handbook of phonetic sciences* (pp. 521-602). Oxford: Wiley–Blackwell.

Fletcher, A. R., & McAuliffe, M. J. (2015). The relationship between speech segment duration and vowel centralization in a group of older speakers. *Journal of the Acoustical Society of America, 138,* 2132-2148.

Fougeron, C., & Keating, P. A. (1997). Articulatory strengthening at edges of prosodicdomains. *Journal of the Acoustical Society of America, 101,* 3728-3740.

Frota, S. (2016). Surface and structure: Transcribing intonation within and across languages. *Laboratory Phonology: Journal of the Association for Laboratory Phonology, 7,* 1-19.

Frota, S., D'Imperio, M., Elordieta, G., Prieto, P., & Vigário, M. (2007). The phonetics and phonology of intonational phrasing in Romance. In P. Prieto (Ed.), *Segmental and prosodic issues in Romance phonology* (pp. 131-154). John Benjamins, Amsterdam.

Gerken, L. (1996). Prosodic structure in young children's language production. *Language, 72,* 683-712.

Gordon, M., & Munro, P. (2007). A phonetic study of final vowel lengthening in Chickasaw. *International Journal of American Linguistics, 73,* 293-330. http://www.jstor.org/stable/10.1086/521729

Gósy, M. (2012). BEA – A multifunctional Hungarian spoken language database. *The Phonetician, 105/106,* 50-61.

Gósy, M. (2017). Frázisvégi nyúlás a spontán beszédben: a fonológiai hosszúság tükröződése [Phrase-final lengthening in spontaneous speech: The reflection of phonological length]. In M. Gósy, & V. Krepsz (Eds., 2017b), *Morfémák időzítési mintázatai a beszédben* [*Temporal patterns of morphemes in speech*] (pp. 134-155). Budapest: MTA Nyelvtudományi Intézet.

Gósy, M., & Krepsz V. (2017a). Magánhangzók nyúlása: mondatpozíció és szóhossz [Lengthening of vowels: Sentence position and word length]. In M. Gósy, & V. Krepsz (Eds.) (2017b), *Morfémák időzítési mintázatai a beszédben* [*Temporal patterns of morphemes in speech*] (pp. 107-133). Budapest: MTA Nyelvtudományi Intézet.

Gósy, M., & Krepsz, V. (Eds., 2017b). *Morfémák időzítési mintázatai a beszédben* [Temporal patterns of morphemes in speech]. Budapest: MTA Nyelvtudományi Intézet.

Hammond, G. R. (Ed., 1990). *Cerebral control of speech and limb movements*. Amsterdam: Elsevier.

Hansson, P. (2003). *Prosodic phrasing in spontaneous Swedish*. Lund: Lund University.

Hockey, B. A., & Fagyal, Zs. (1999). Phonemic length and pre-boundary lengthening: an experimental investigation on the use of durational cues in Hungarian. In *Proceedings of the 14th International Congress of Phonetic Sciences* (pp. 313-316). San Francisco.

Hofhuis, E., Gussenhoven, C., & Rietveld, A. (1995). Final lengthening at prosodic boundaries in Dutch. In *Proceedings of the ICPhS 1995*. Vol 1 (pp. 154-157). Stockholm.

Hooper, C. R., & Cralidis, A. (2009). Normal changes in the speech of older adults: You've still got what it takes; it just takes a little longer! *Perspectives on Gerontology, 14,* 47-56.

Huber, J. E. (2008). Effects of utterance length and vocal loudness on speech breathing in older adults. *Respiratory Physiology and Neurobiology, 164,* 323-330.

Jacewicz, E., Fox, R. A., & Wei, L. (2010). Between-speaker and within-speaker variation in speech tempo of American English. *Journal of the Acoustical Society of America*, *128,* 839-850.

Jong de, K., & Zawaydeh, A. B. (1999). Stress, duration, and intonation in Arabic word-level prosody. *Journal of Phonetics, 27,* 3-22.

Kachkovskaia, T. (2014). Phrase-final lengthening in Russian: Pre-boundary or pre-pausal? In A. Ronzhin, R. Potapova, & V. Delic (Eds.), *Speech and computer* (pp. 353-359). Novi Sad: Springer International Publishing.

Kail, R., & Salthouse, T. A. (1994). Processing speech as a mental capacity. *Acta Psychologica, 86,* 199-225.

Kassai, I. (1982). A magyar beszéd időtartamviszonyai [Temporal patterns of Hungarian speech]. In K. Bolla (Ed.), *Fejezetek a magyar leíró hangtanból* [*Chapters from a descriptive phonetics of Hungarian*] (pp. 115-154). Budapest: Akadémiai Kiadó.

Kent, R. D. (2000). Research on speech motor control and its disorders: A review and prospective. *Journal of Communication Disorders, 33,* 391-428.

Klatt, D. H. (1975). Vowel lengthening is syntactically determined in a connected discourse. *Journal of Phonetics, 3,* 129-140.

Kohler, K. J. (1983). Prosodic boundary signals in German. *Phonetica, 40,* 89-134.

Koponen, E. & Lacerda, F. (2003). Final lengthening in infant directed speech may function as a cue to phrase constituents. *PHONUM, 9,* 9-12.

Krepsz, V. (2017). Szótag nyúlása a frázisvégen [The lengthening of phrase-final syllables]. In M. Gósy, & V. Krepsz (Eds., 2017b), *Morfémák időzítési mintázatai a beszédben* [*Temporal patterns of morphemes in speech*] (pp. 156-174). Budapest: MTA Nyelvtudományi Intézet.

Krull, D. (1997). Prepausal lengthening in Estonian: Evidence from conversational speech. In I. Lehiste, & J. Ross (Eds.), *Estonian prosody: Papers from a symposium* (pp. 136-148). Tallinn: Institute of Estonian Language.

Lee, T-L., He, Y-F., Huang, Y-J., Tseng, S-Ch., & Eklund, R. (2004). Prolongation in spontaneous Mandarin. In *Proceedings of the 8th International Conference on Spoken Language Processing* (pp. 2181-2184). Jeju Island, Korea.

Lieshout, P. H. H. M. van, Starkweather, C. W., Hulstijn, W., & Peters, H. F. M. (1995). Effects of linguistic correlates of stuttering on EMG activity in nonstuttering speakers. *Journal of Speech and Hearing Research, 38,* 360-372.

Lindblom, B. (1968). Temporal organization of syllable production. In *Speech Transmission Laboratory Quarterly Progress* 9, (pp. 1-6). Stockholm: Royal Institute of Technology.

Liss, J. M., Weismer, G., & Rosenbek, J. C. (1990). Selected acoustic characteristics of speech production in very old males. *Journal of Gerontology, 45,* 35-45.

Maastricht, L. van, Krahmer, E., Swerts, M., & Prieto, P. (2016). *Learning L2 rhythm.* Paper presented at Speech Prosody 2016, Boston, United States.

Markó, A., & Kohári, A. (2015). Glottalization and timing at utterance final position in Hungarian: Reading aloud vs. spontaneous speech. In *Proceedings of the 18th International Congress of Phonetic Sciences,* paper 0722.

Nagano-Madsen, Y. (1992). *Mora and prosodic coordination. A phonetic study of Japanese, Eskimo and Yoruba.* PhD thesis. Kent/Lund University.

Nakai, S., Kunnari, S., Turk, A., Suomi, K., & Ylitalo, R. (2009). Utterance-final lengthening and quantity in Northern Finnish. *Journal of Phonetics, 39,* 29-45.

Oller, K. D. 1973. The effect of position in utterance on speech segment duration in English. *Journal of the Acoustical Society of America, 54,* 1235-1247.

Plüschke, M., & Harrington, J. (2013). The domain of phrase final lengthening in Estonian. In E. L. Asu, & P. Lippus (Eds), *Proceedings of International Conference: Nordic Prosody XI* (pp. 293-302). Tartu. Frankfurt am Main: Peter Lang Verlag.

Rao, R. (2010). Final lengthening and pause duration in three dialects of Spanish. In M. Ortega-Llebaria (Ed.), *Proceedings of the 4th Conference on Laboratory Approaches to Spanish Phonology* (pp. 69-82). Somerville, MA: Cascadilla Proceedings Project.

Rodríguez-Aranda, C., & Jakobsen, M. (2011). Differential contribution of cognitive and psychomotor functions to the age-related slowing of speech production. *Journal of the International Neuropsychological Society, 17,* 1-15.

Shipp, T., Qi, Y., Huntley, R., & Hollien, H. (1992). Acoustic and temporal correlates of perceived age. *Journal of Voice, 6,* 211-216.

Siptár, P., & Törkenczy, M. (2000). *The Phonology of Hungarian*. Oxford: Oxford University Press.

Smith, B. L., Wasowicz, J., & Preston, J. (1987). Temporal characteristics of the speech of normal elderly adults. *Journal of Speech and Hearing Research, 30,* 522-529.

Snow, D. (1994). Phrase-final syllable lengthening and intonation in early child speech. *Journal of Speech and Hearing Research, 37,* 831-840.

Sweeting, P. M., & Baken, R J. (1982). Voice onset time in normal-aged population. *Journal of Speech and Hearing Research, 25*(1), 129-134.

Torre, P., & Barlow, J. A. (2009). Age-related changes in acoustic characteristics of adult speech. *Journal of Communication Disorders, 42,* 324-333.

Turk, A. (2007). Multiple targets of phrase-final lengthening in American English words. *Journal of Phonetics, 35,* 445-472.

Turk, A., & Shattuck-Hufnagel, S. (2007). Multiple targets of phrase-final lengthening in American English words. *Journal of Phonetics, 35,* 445-461.

Weismer, G., & Liss, J. M. (1991). Speech motor control and aging. In D. N. Ripich (Ed.), *Handbook of geriatric communication disorders* (pp. 205-225). Austin: Pro-ed Press.

White, L. S. (2002). *English speech timing: A domain and locus approach.* Ph.D. dissertation. University of Edinburgh.

White, L., & Mády, K. (2008). The long and the short and the final: Phonological vowel length and prosodic timing in Hungarian. In P. A. Barbosa, S. Madureira, & C. Reis (Eds.), *Proceedings of the Speech Prosody 2008 Conference* (pp. 363-367). Campinas, Brazil.

Wightman, C. W., Shattuck-Hufnagel, S., Ostendorf, M., & Price, P. J. (1992). Segmental durations in the vicinity of prosodic phrase boundaries. *Journal of the Acoustical Society of America, 91*(3), 1707-1717.

Winkler, R., Brückl, M., & Sendlmeier, W. (2003). The aging voice: An acoustic, electroglottographic and perceptive analysis of male and female voices. In *Proceedings of the ICPHS 2003* (pp. 2869-2872).

Winkworth, A. L., Davis, P. J., Adams, R. D., & Ellis, E. (1995). Breathing patterns during spontaneous speech. *Journal of Speech and Hearing Research, 38,* 124-144.

Wohlert, A.B., & Smith, A. (1998). Spatiotemporal stability of lip movements in older adult speakers. *Journal of Speech, Language, and Hearing Research, 41,* 41-50.

Xue, S. A., & Hao, G. J. (2003). Changes in the human vocal tract due to aging and the acoustic correlates of speech production: A pilot study. *Journal of Speech, Language and Hearing Research, 46,* 689-701.

Yorkston, K. M., Bourgeois, M. D., & Baylor, C. R. (2010). Communication and aging. *Physical Medicine and Rehabilitation Clinics of North America, 21*(2), 309-319.

Zraick, R. I., Gregg, B. A., & Whitehouse, E. L. (2006). Speech and voice characteristics of geriatric speakers: A review of the literature and a call for research and training. *Journal of Medical Speech and Language Pathology, 14,* 133-142.

LARYNX MOVEMENT IN THE PRODUCTION OF GEORGIAN EJECTIVE SOUNDS

Alexandra BÜCKINS, Reinhold GREISBACH, & Anne HERMES
IfL-Phonetics, University of Cologne, Germany
a103690@smail.uni-koeln.de, reinhold.greisbach@uni-koeln.de,
anne.hermes@uni-koeln.de

Abstract
In this study, we present a non-invasive method for investigating laryngeal movement in the production of ejective sounds. Being non-invasive, this method can be used easily in the study of spontaneous speech. Typically, EMA is used to track the tongue and lip movements in speech production. In this study, we recorded four Georgian native speakers with four sensor coils on the outside of the skin – just above the larynx – in the area of the cricoid cartilage. The analysis reveals that there is considerably greater movement of the coils during the production of ejectives as compared to pulmonal sounds. These movement patterns of the skin above the larynx are admittedly of very complex nature. To attribute the movement solely to the larynx is problematic. Nonetheless, this method may help to understand the production mechanism of ejectives.

Keywords: Georgian, Ejective, Larynxmovement, EMA

1 Introduction

Ejective sounds are relatively rare in European languages and occur only in the Caucasus region (e.g., in Georgian), and by implication are not very well investigated. Ladefoged and Maddieson (1996) refer to ejectives as "not at all unusual sounds, occurring in about 18 percent of the languages of the world", but in quite diverse language families (e.g., Mayan and Chadic). The ejective production mechanism can be applied to produce plosives, affricates and fricatives both midsagitally and laterally. Plosive and affricate ejectives are most common, while fricative and lateral ejectives are only found in a handful of languages worldwide (Maddieson, 2013). Velar articulations seem to be most favored for ejective stops, cf., Greenberg (1970) and Maddieson (1984). Uvular ejective stops are also reported to be fairly common. Amongst the affricates, [ʦ'] and [ʧ'] seem to be widely spread (Maddieson, 2013).

In the Caucasus almost every language of the indigenous language families exhibits ejectives, i.e., languages from the Kartvelian (Southwest Caucasian), from the Nakho-Dagestanian (Northeast Caucasian) and from the Abkhazo-

DOI: http://doi.org/10.18135/CAPSS.127

Adyghean (Northwest Caucasian) language families. But there are also Indo-European languages spoken in the area such as Ossetic and East Armenian which have ejectives included into their phoneme system. Thus, the presence of ejectives seems to be an areal phenomenon of the Caucasus.

In these languages as well as in the languages of the indigenous Caucasian language families there is typically a threefold opposition between voiced, voiceless and ejective plosives and affricates. The plosive triples are most common for the labial, dental, velar and uvular places of articulation, while affricate triples are typically alveolar or palato-alveolar. For the East Caucasian languages, we typically find an additional binary opposition of the lateral ejectives (vs. their voiceless counterparts). Some Northwest Caucasian languages (e.g., Kabardian, Adyghe) also have a threefold opposition for fricatives, which include ejective fricatives (Klimov, 1994; Vinogradov, 1967).

In Georgian, there are three-way oppositions for labial [b ~ p ~ p'], dental [d ~ t ~ t'] and velar [g ~ k ~ k'] plosives on the one hand, and alveolar [dz ~ ts ~ ts'] and palato-alveolar [dʒ ~ tʃ ~ tʃ'] affricates on the other. Additionally, we find a singleton uvular ejective, mostly denoted [q'], which may phonetically surface, depending inter alia on speaker and speaking style, as an ejective plosive, fricative or affricate.

The production of ejectives involves a non-pulmonal airstream mechanism. The airstream is invoked by raising of the closed larynx. At the same time there has to be a constriction (plosive, fricative, or affricate) taking place in the supraglottal space, namely the mouth. The raising of the closed glottis leads to an increase in pressure in the space behind the constriction. Due to greater pressure drop, ejectives sound more prominent compared to pulmonal sounds (Ladefoged & Maddieson, 1996).

Figure 1 illustrates the phases of ejective plosive production. Phase one represents oral and glottal closure, and raising of the larynx. Phase two indicates the compression of air inside the enclosed oral section. Finally, phase three points out the oral release burst, while the glottis remains closed.

Figure 2 illustrates the main difference between ejective plosive production and ejective affricate production. In phase three the larynx is lifted up even further, while the larynx remains closed. The oral release burst is accompanied by friction.

The following example out of our data represents a typical acoustic pattern of an ejective plosive (Figure 3), in this case for the Georgian syllable [p'a]. This pattern is characterized by three acoustic phases:

- aperiodic noise of pressure release (ASP)
- silence (PAUS)
- (optional) creaky transition into vowel (CREAK).

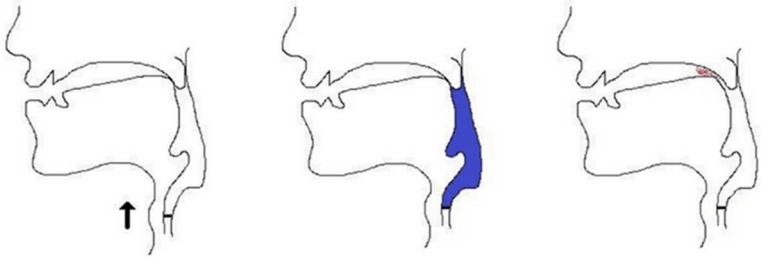


Figure 1.
Phases of ejective plosive production

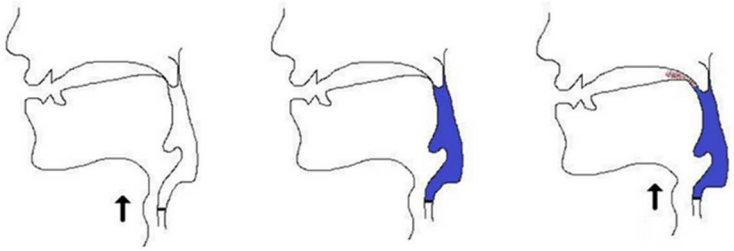


Figure 2.
Phases of ejective affricate production

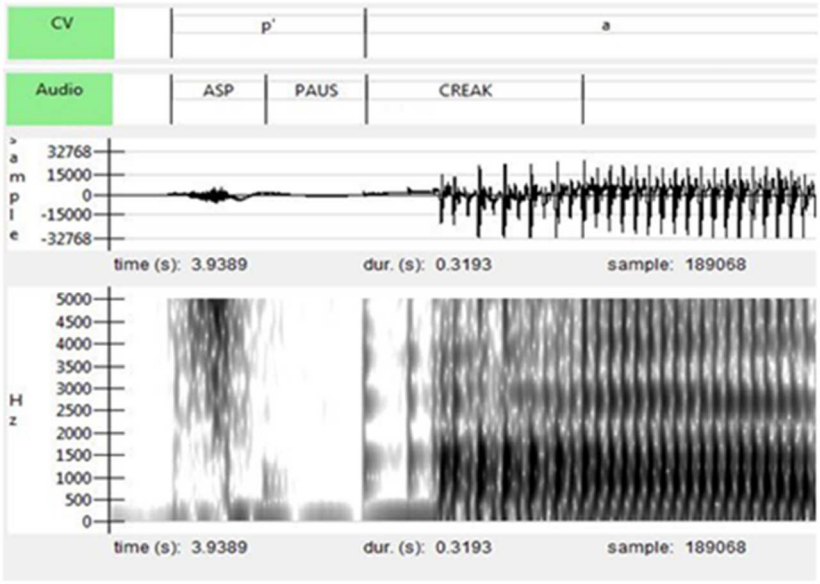


Figure 3.
The acoustics of an ejective plosive; segments in the CV tier; sound phases (Audio tier); oscillogram and sonogram for syllable [p'a], speaker I (f)

Georgian specifically distinguishes on the phonetic level between voiceless plosives / affricates, strong aspirated plosives / affricates, and ejective plosives / affricates. In the traditional grammar of Georgian they are often called voiced plosives / affricates, voiceless plosives / affricates and glottalized plosives / affricates, respectively (e.g., Tschenkeli, 1958, p. XLVII; Cherchi, 1999, p. 2).

Not much is known about variation in ejective production, be it inter-language or inter-speaker specific. Grawunder et al. (2010) give an overview on the production of ejectives in a number of Caucasian languages, but they only focus on specific phonetic parameters not comparing the production across languages. Listening to news in radio broadcasts leads subjectively to the impression that e.g., the ejective production in Avar (a Nakho-Dagestanian language) is stronger, causing auditorily a click-like impression of Avar ejectives compared e.g., to Georgian ejectives, which sound smoother.

Empirical evidence for variation in ejective production can be found in Lindau (1984), who points out significant cross-linguistic and inter-speaker variation in the comparison of velar ejectives in Hausa and Navajo.

Independent from speech rate, Lindau (1984) proposes the main difference between Navajo and Hausa speakers to be the long glottal closure in Navajo. The Navajo glottal closure is furthermore released into creaky voice.

Due to the small number of speakers typical for studies on ejective sounds it is difficult to decide whether certain parameters of ejective production are cross-linguistic differences or speaker-specific phenomena.

Articulatory investigations on ejectives are relatively rare. Grawunder et al. (2010) describe in an impressionistic way the elevation of the larynx during ejective production for one speaker of Georgian.

Alongside the examination of prosodic features or the consequences of speech disfluency (e.g., repetitions, repairs, pauses) spontaneous speech offers the most intuitive and therefore natural data of articulatory gestures. To visualize articulatory movement in a spontaneous speech setting with no restriction on speech aside from topic or task given by the supervisor is problematic due to the necessary invasive methods. One of the most common methods to investigate articulatory movement of the frontal area of the vocal tract is Electromagnetic Articulography (EMA). Sensory connector coils are positioned on the lips and inside the subject's mouth to display exact information on lip, tongue and jaw movement.

In this pilot study, we propose the possibility to record articulatory data of laryngeal mechanisms in spontaneous speech. Thus, EMA is used as a non-invasive method to analyze larynx movement in the production of Georgian ejective sounds. This method is non-invasive in that it is quick to adjust, and comfortable for the speaker, in that it does not restrain articulation in any way.

Our main motivation was to

- visualize the ejective production mechanism
- display articulatory movement of the larynx non-invasively
- collect rare articulatory data of ejective sounds

We expect greater movement of the sensory connector coils during the production of ejectives as compared to pulmonal sounds, as well as noticeable changes in the data. Therefore, we propose that the skin movement pattern for ejectives will be more prominent than the movement pattern for pulmonal sounds with the same place and manner of articulation. Furthermore, the skin movement pattern for ejective affricates is expected to be more prominent than the movement pattern for ejective plosives due to the larger raising of the larynx. As male and female speakers differ in gender-specific laryngeal anatomy, more conclusive data might be observed in male speakers.

2 Method

Usually, one can observe the larynx movements in male speakers very easily. Due to their naturally prominent anatomy of the thyroid cartilage the raising and lowering of the larynx is visible. This could be recorded on video, but would *inter alia* require the speaker to be clean-shaven. An adaptive use of the Electromagnetic Articulography (EMA) avoids the necessity of prominent larynx anatomy and the exclusion of female speakers from the study.

Typically, EMA is used to monitor tongue and lip movements in speech production. EMA requires coils behind the ears, the bridge of the nose, and on tongue root, body and tip, which were also included in this study. Figure 4 illustrates how we placed four additional sensor coils on the outside of the skin just above the larynx in the area of the cricoid cartilage of the speaker. A further coil was added on the back of the neck to control for head rotation.

Figure 5 serves as an example of the EMA coils attached to record the movement of the larynx. Recordings were done with the AG501 EMA (Carstens Medizinelektronik) with a sampling rate of 250 Hz. EMA allows to monitor the position of up to 16 coils in a magnetic field, which is positioned above the speaker. The audio signal was registered synchronously at 48 kHz. Labeling was done within EMU Speech Database System (Cassidy & Harrington, 2001), further analysis and graphs with the R programming language.

To focus on the differences between the ejective sounds and their pulmonal counterparts we segmented every word into its segments and evaluated the position of the laryngeal coils at the temporal midpoints. As the movement of the head is expected to be small from one word to the next one within a recording sweep, but may be great from one sweep to the next, we averaged the coil position of the 3 plosives / affricates (taken in the center of every friction

part) in every sweep of e.g., [ts'-], [dz-] and [ts-] and set this average to be the origin of the coordinate system (i.e., the data was ipsativated).

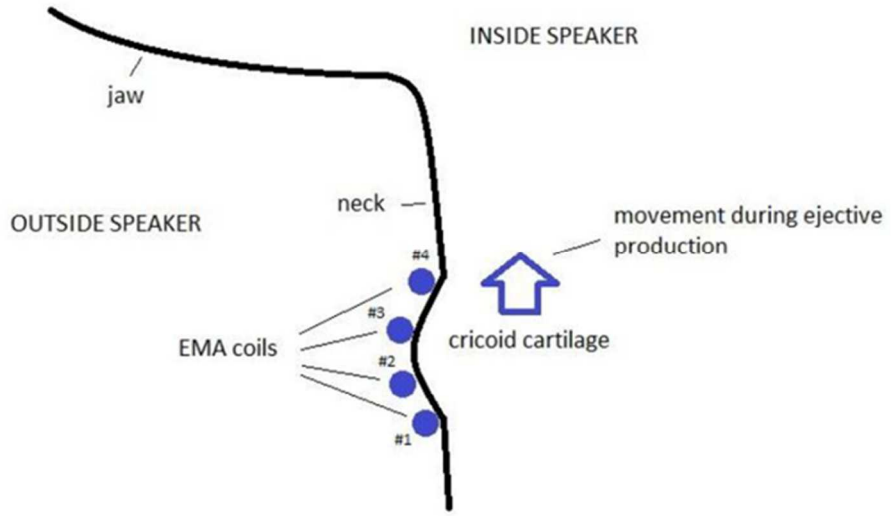


Figure 4.
Position of the EMA coils on the outside of the neck just above the larynx in midsagittal plane

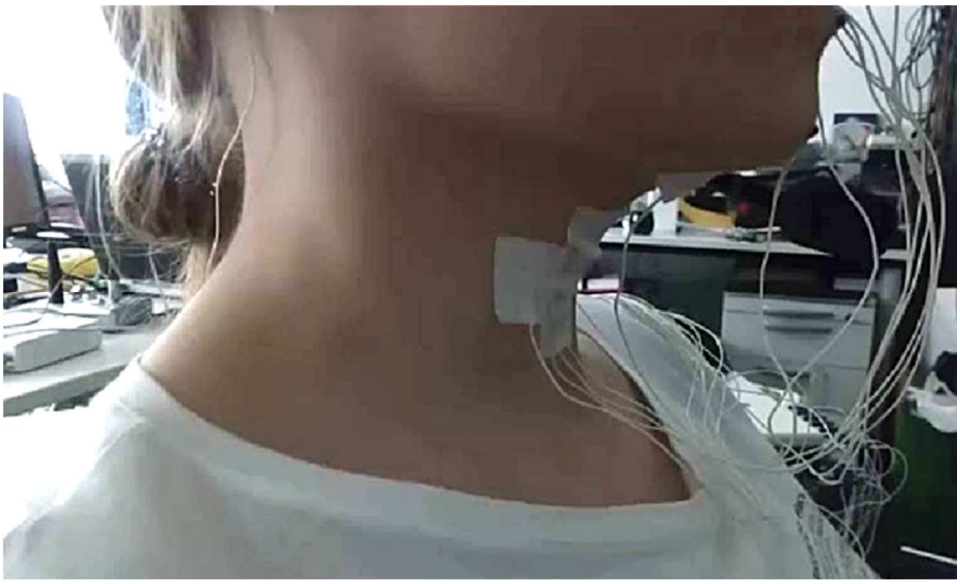


Figure 5.
EMA coils for speaker E during EMA recording

2.1 Subjects

For this study, four native Georgian speakers between 20 and 40 years were recorded (three female, one male). The female speakers are referred to as E, I and N. The male speaker is referred to as B. All speakers lived in Cologne or proximate vicinity at the time of recordings. They were all fluent in their native language and also competent in its literacy.

2.2 Speech Material

The corpus of this pilot study is based on the ejective contrast in Georgian, which is summarized for the apical consonants in Table 1.

For this pilot study we used word lists, consisting of minimal pairs or triples. The contrastive pairs and triples are summarized in Table 2. To avoid the influence of strong prosodic boundaries (end of the utterance), the first word of the respective contrast was repeated at the end of every recorded sweep and excluded from the analysis.

Table 1. Exemplary plosive and affricate contrasts in Georgian

Plosive	**Affricate**	**Grammar**	**Phonetics**
[d]	[dz]	voiced	voiceless
[t]	[ts]	voiceless	strong aspirated
[t’]	[ts’]	glottalized	ejective

Table 2. Contrastive minimal pairs and triples

პაპა [p’ap’a] ‘grandfather’	ფაფა [papa] ‘porridge’	
	ფარი [pari] ‘shield’	ბარი [bari] ‘spade’
პური [p’uri] ‘bread’	ფური [puri] ‘cow’	
ტარი [t’ari] ‘handle’	თარი [tari] ‘tar’ (musical instrument)	დარი [dari] ‘good weather’
წელი [ts’eli] ‘year’	ცელი [tseli] ‘scythe’	ძელი [dzeli] ‘log’
წერა [ts’era] ‘write’	ცერა [tsera] ‘little finger’	ძერა [dzera] ‘vulture’
კერა [k’era] ‘hearth’	ქერა [kera] ‘blond’	

Seven repetitions of these contrasts, and therefore 476 target words were taken into analysis. The subjects simply read them out.

3 Results

In the following, we present the results for the movement of the coil fixed above the larynx. Since all of the coils attached on the larynx did show a similar movement, we focused on the upper central coil #3 (see Figure 4).

The movement (of the skin) above the larynx can be illustrated by discrete production phases, as displayed in Figure 6 for [ts'eli]. It is possible to stepwise track the movement pattern of the coil centered above the larynx. The trace of the coil starts on the bottom left. It represents the closure phase (here the larynx is lowered and in the back). The following part (denoted /ts'/) shows the phase of the non-pulmonal burst mechanism: the closed glottis is pushed up, thus increasing the air pressure above the glottis and producing a fricative noise in the alveolar region. When the upward movement stops, the fricative noise in the audio signal ends immediately and the vowel (denoted /e/) begins. Throughout the rest of the word (denoted /li/) the laryngeal coil moves continuously downwards.

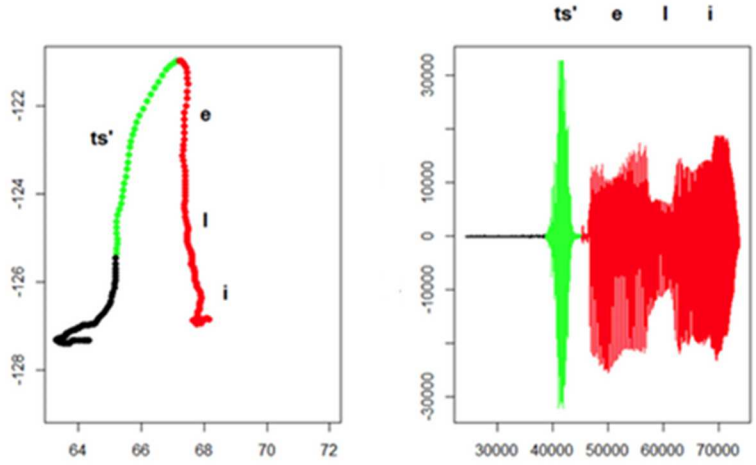


Figure 6.
Dynamics of coil centered above the larynx (left; scales in mm) and acoustic signal (right) for [ts'eli] relative to the bite plane; speaker N (f)

3.1 Ejective affricate contrasts

We find a relatively consistent pattern of the skin position just above the larynx over all 7 repetitions. The following example refers to the affricate contrast in the minimal triple [ts'eli] vs. [dzeli] vs. [tseli] and includes all 4 speakers (Figure 8). Displayed is the coil position in the center of every sound relative to the x-y-coordinate system defined by the bite plane. Depending on the general physiological posture of a person (e.g., upright or buckled) and the posture of the

head during the recording the contour of the neck as well as the movement direction of the larynx will be more or less perpendicular to this coordinate system. Any componential interpretation of the patterns in upward-downwards and forward-backward direction has therefore to be done with caution.

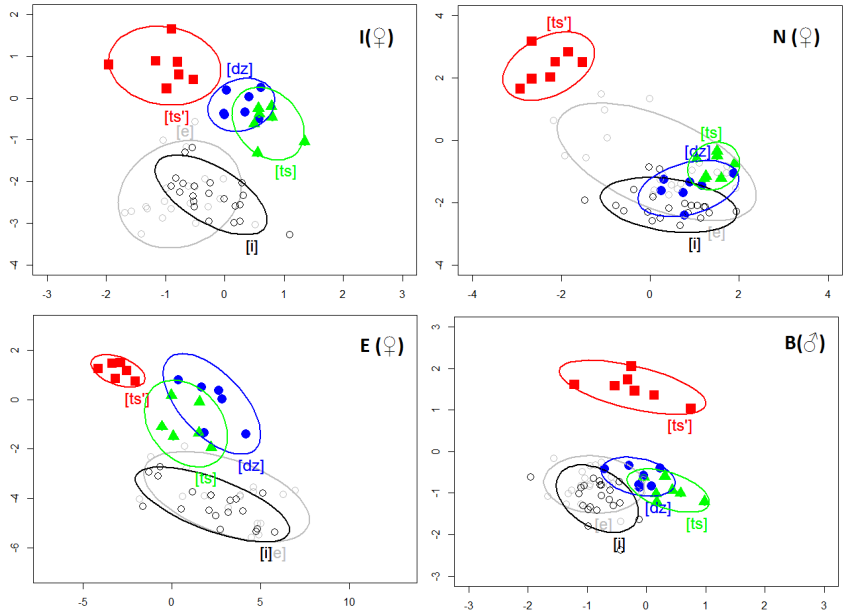


Figure 8.
Larynx coil in [ts'eli] vs. [dzeli] vs. [tseli] for speakers I, N, E (all female) and B (male); scales are in mm; the ellipses denote the 80% confidence interval

The 80% confidence ellipses surrounding the squares (Figure 8) reflect the position of the larynx coil while producing the affricate ejective [ts']. It is positioned above all other ellipses especially the one surrounding the triangles (voiceless [ts]) and the filled circles (voiced [dz]), which indicates the larynx has been in a more raised position. For the vowels [e] (grey empty circles) and [i] (black empty circles) of the minimal triple frame ([-eli]) the larynx is in a lower position for most speakers as compared to all of the consonants. These coordinates serve as points of reference. Furthermore, the ejective coordinates (squares) are displayed on the left, for all three female speakers, which means that the skin is being pulled back in relation to the vowels, not pushed out. For the male speaker there is no backward shift in relation to the reference bite plane.

3.2 Ejective plosive contrasts

The following example (Figure 9) refers to the ejective plosive contrasts in the triple [t'ari] vs. [dari] vs. [tari] and includes 7 repetitions of all 4 speakers. The

results are comparable to the ejective affricate contrasts (Figure 8), even though the movement in up-down-direction is not as prominent for the ejective plosives.

To compensate for any movements of the speaker during a recording session, we took the ipsative coordinate values of every minimal triple taken in the center of every consonant. What we do compare are the differences between the consonants uttered in immediate neighborhood of an utterance.

Thus, a direct comparison between plosive ejectives and affricate ejectives is not possible, because the reference points of the coordinate systems depend on the coordinates of all members of the triple the sound in question belongs to, and not on the absolute values. All we can say is that the distribution patterns between affricates and plosives are similar and the magnitude of the displacement relative to the vowels is larger for ejective affricates than for ejective plosives.

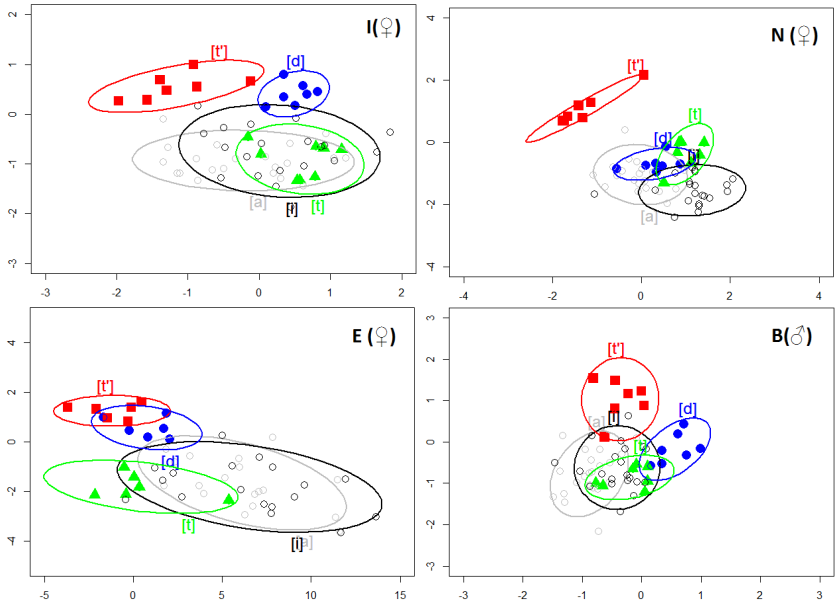


Figure 9.
Larynx coil in [t'ari] vs. [dari] vs. [tari] for speakers I, N, E (all female) and B (male); scales are in mm; the ellipses denote the 80% confidence interval

4 Summary

This pilot study on Georgian provides preliminary evidence that there is a larger movement of the skin in ejectives compared to pulmonal sounds. The skin above the larynx is shifted higher upwards than in the following vowels. The magnitude of the skin movements depends on the individual anatomical disposition of the speaker, but seems to be a little bit higher for affricates than

for plosives. No difference in magnitude of the movement between male and female speakers was found.

In the female speakers there is a simultaneous backward shift of the skin, probably due to the angle between neck and bite plane. For the male speaker no backward shift is noted. This difference might be attributed to the anatomical differences of the male and female cricoid cartilage. Fixed on the skin slightly above the cricoid cartilage in rest position, the coil under observation is moving upwards, as the skin is pulled upwards by the cricoid cartilage. Typically, the cricoid cartilage is moving further up than the skin above it, which is clearly visible in male speakers. Thus, in the most upward position the coil will no longer be situated above the cricoid cartilage, but probably exactly on the top of the cricoid cartilage or even below it. As in male speakers the cricoid cartilage forms an outward bulb on the skin, the coil is expected to be shifted in the front-back direction as well, either forward or backward, depending on its initial position.

5 Conclusions

We were able to visualize movement patterns of the ejective production mechanism, but we are aware that these patterns are of a very complex nature. The movement of the skin above the larynx is influenced not only by the movement of the larynx, but also by

- movement of the hyoid bone and attached muscles on the neck while speaking,
- movement of the mimic muscles while speaking and opening the jaw (platysma),
- movement of the head and attached muscles on the neck while speaking (sternomastoid) (Gray & Drake, 2008).

To attribute the recorded movements exclusively to the larynx is therefore problematic, as well as controlling for all those factors above. Nonetheless we did observe movements that are consistent with the laryngeal mechanism for ejective production, and propose our adapted version of Electromagnetic Articulography to be used for laryngeal research in spontaneous speech.

References

Cassidy, S., & Harrington, J. (2001). Multi-level annotation in the Emu speech database management system. *Speech Communication, 33,* 61-77.

Cherchi, M. (1999). *Georgian.* München, Newcastle: Lincom Europa.

Gray, H., & Drake, R. L. (2008). *Gray's atlas of anatomy.* Philadelphia: Churchill Livingstone.

Grawunder, S., Simpson, A., & Khalilov, M. (2010). Phonetic characteristics of ejectives – samples from Caucasian languages. In S. Fuchs, M. Toda, & M. Zygis (Eds.), *Turbulent sounds* (pp. 209-244). Berlin: De Gruyter Mouton.

Greenberg, J. (1970). Some Generalizations Concerning Glottalic Consonants, Especially Implosives. *International Journal of American Linguistics, 36,* 123-145.

Klimov, G. A. (1994). *Einführung in die kaukasische Sprachwissenschaft.* Hamburg: Buske.

Ladefoged, P., & Maddieson, I. (1996). *The sounds of the World's Languages.* Oxford: Blackwell.

Lindau, M. (1984). Phonetic differences in glottalic consonants. *Journal of Phonetics, 12,* 147-155.

Maddieson, I. (1984). *Patterns of Sounds.* Cambridge: Cambridge University Press.

Maddieson, I. (2013). Glottalized Consonants. In M. S. Dryer, & M. Haspelmath (Eds.), *The World Atlas of Language Structures Online.* Leipzig: Max Planck Institute for Evolutionary Anthropology. (Available online at http://wals.info/chapter/1, Retrieved on 04.05.2018.)

Tschenkéli, K. (1958). *Einführung in die Georgische Sprache.* Zürich: Amirani Verlag.

Vinogradov, V. V. (1967). *Jazyki narodov SSSR.* Moskva: Ibersijsko-Kavkazskie jazyki.

COVERT CONTRAST IN THE EARLY DEVELOPMENT OF SPEECH SOUNDS IN CHILDREN USING COCHLEAR IMPLANTS

Ruth HUNTLEY BAHR[1], Terry GIER[2], Laura CONOVER[1]
[1]University of South Florida, Tampa, FL USA,
[2]Our Children's Academy of Winter Haven, Winter Haven, FL USA
rbahr@usf.edu, terry2@mail.usf.edu, lconover1@mail.usf.edu

Abstract
Covert contrasts represent intermediate productions that allow us broader insight into how children acquire a phonological system. However, little is known about the use of covert contrast in the development of speech sounds in children with cochlear implants (CIs). In particular, are these children using covert contrast in the same way that children with normal hearing (NH) do? Nine congenitally deafened children with CIs, ages 2;11 to 6;4 years (M = 4;9), who were implanted before age 3 were matched to typically developing children by articulation ability and gender. Their VCV productions from the OlimSpac were rated by 33 experienced listeners on an equal appearing interval scale to rate the phonetic accuracy of /t/ and its production as a substitution for /d/ and /ʧ/. Results indicated no differences in [t] production across groups. However, children with NH had a large, well-developed contrast between /t/ and /d/, but the later developing /tʃ/ showed little contrast with /t/. Children with CIs demonstrated the opposite trend. Their [t] for /d/ substitutions were more /t/-like, suggesting insufficient covert contrast for the voicing difference between these phones. However, they displayed a larger contrast for /t/ and /tʃ/ than the children with NH.

Keywords: speech production, speech sound development, cochlear implants, covert contrast, listener ratings

1 Introduction

Speech perception may be categorical, but speech production is not (Munson et al., 2010). Since human speech is characterized by inter- and intraspeaker variation, the precision of speech sounds can fall anywhere along the continuum between two phonemically similar sounds. This phenomenon rarely presents a problem in normal conversation, as listeners are biased to resolve ambiguous sounds to perceive words rather than nonwords (Strombergsson, Salvi, & House, 2015). However, a listener's strong sense of categorical perception begins to break down when he/she is asked to estimate the "goodness" of a sound production on a scale between two potential target phonemes, rather than using a simple forced-choice task (Strombergsson et al., 2015). While not always the

DOI: http://doi.org/10.18135/CAPSS.139

case, some of these ambiguous or less "good-fitting" sounds can be acoustically different from the target phonemes in ways that are detectible using spectrograms and other acoustic methods. As such, these ambiguous productions comprise a category of sounds known as covert contrasts, which are defined as "impressionistically homophonous [speech sound] categories that can be reliably distinguished at the phonetic level" (Kirby, 2011, p.1090).

In this paper, we will focus on the use of covert contrast in the development of speech sounds by children with cochlear implants (CIs) as one of several factors that influence speech development. Since each language has its own phonological contrasts, children must master specific perceptual and articulatory skills to become a proficient speaker of that language. In particular, infants need to learn which aspects of a sound function as unique cues for the production of meaningful speech. As young children develop phonetic categories, they go through a stage where their productions may be perceptually unreliable but acoustically distinct (i.e., covert contrasts). This type of production is often used as evidence against the traditional view that pronunciation shifts in children are caused by solely phonological changes (Scobbie et al., 1996; Strombergsson et al., 2015). As such, covert contrasts represent intermediate productions that are their own stage of learning, allowing researchers broader insight into how children acquire a phonological system (Hewlett & Waters 2004; Munson et al., 2012; Munson et al., 2017). Additional support for the use of covert contrast in speech development comes from the fact that children with speech sound disorders who produce covert contrasts have much better prognoses in treatment than those who do not (Byun et al., 2016).

1.1 Covert contrast

The fact that a child produces a covert contrast between two phonemes suggests that he/she can perceive some difference between them (Byun et al., 2016), which can then be refined into distinct phonemes. Research has shown that covert contrast has been observed for place of articulation for stops (Forrest et al., 1990), place of articulation for fricatives (Li et al., 2009), and voicing for stops (Macken & Barton, 1980). Of particular interest here is the research that describes factors that influence the perception of covert contrast. For instance, the context of the speech sound influences its perception. Sounds presented in real words are easier to detect than those presented in non-words (as reviewed by Strombergsson et al., 2015). These researchers also noted that the frequency of the phonotactic context and the listener's level of experience also influences the perception of covert contrast. Finally, Munson et al. (2010) demonstrated that speaker age influences the perception of covert contrast. These researchers described how older speakers' ambiguous phoneme productions were more

likely to be judged as errors because the listener expected that elders should have well-developed phonemes.

The perception of subtle differences in speech sounds is essential for individuals who may receive a distorted or diminished speech signal, such as children who use cochlear implants (CIs). It has been well-documented that the signal delivered by CIs, although adequate for reasonably accurate speech perception, is significantly degraded in relation to the acoustic information that is available to a person with normal hearing. This is due to the processing methods common to CIs (Pisoni et al., 1999; Spencer, 2002). This modified speech signal might influence the speech features noted in the speech production of children who use CIs and might negatively impact the production of covert contrast.

1.2 Speech production in children with CI

There is significant variation in the speech production skills of children who use CIs. One of the predictors of speech accuracy is whether or not the child has successfully formed phonological representations of the speech sounds they are attempting to use (James et al., 2008). The production of covert contrast indicates that such phonological representations are developing, as the speaker is producing some acoustic difference between attempts that may be perceived as phonetically similar. Successful use of covert contrasts in children in the early stages of cochlear implant use would suggest they are likely to achieve better speech intelligibility than those who produce clear phoneme substitutions for longer periods of time.

Despite demonstrated variability in speech intelligibility, children with hearing loss show an initial accelerated growth in phoneme development after CI implantation, followed by a plateau where consonantal order of acquisition generally mirrors that of NH children, but at a slower rate (Blamey et al., 2001; Serry & Blamey, 1999; Spencer & Guo, 2013). This finding is more robust when device experience, as opposed to the chronological age of the child, is used as the metric for comparison with typically developing children (Flipsen, 2011). Nevertheless, some studies have suggested that the order of consonant acquisition in children with CIs differs slightly from that of typically developing children. Ertmer et al. (2012) found that some late-developing phonemes were produced more accurately than middle- or early-developing phonemes. Several other studies have identified that the /t/ productions of children with CIs were significantly less accurate than those of children with normal hearing (NH; Blamey et al., 2001, Chin, 2003; Ertmer et al., 2012). These same studies showed that production of /d/ was not similarly delayed. Additionally, the later-developing affricate /ʧ/ has been shown to emerge in children with CI significantly earlier than in children with NH (Ertmer et al., 2012; Spencer & Guo, 2013). Nevertheless, these trends have not been noted consistently, perhaps

due to differences in research methodology, such as whether the researchers used broad or narrow transcription.

Given the noted differences in the speech production ability of children with CIs, it is likely that phonetic transcription alone will not adequately describe their early speech productions. This hypothesis led researchers to consider different measurement techniques. For instance, Schellinger and colleagues (2017) demonstrated that listeners could distinguish small, but statistically significant differences, in phonetic detail in children's speech when asked to rate productions on a visual analogue scale (VAS). Hence, increasing the depth of perceptual choice could produce a tool that can reliably reveal covert contrasts that listeners have been unable to identify using forced-choice or transcription measures alone. Such a finding is useful because assessing the presence of covert contrasts in speech productions holds clinical value (Munson et al., 2012). For instance, children who produce covert contrasts have a much higher likelihood of learning to correctly pronounce target phonemes than those who do not (Strombergsson et al., 2015), and children with speech-sound disorders who do not produce covert contrasts typically require longer treatment times (Tyler et al., 1993). Finally, the ability to measure the presence of covert contrast would imply the ability to track the progress of phoneme development from immaturity to maturity. Clinicians would not be forced into a choice of either correct or incorrect but would be able to track subtle changes during treatment.

1.3 Purpose of the study

As demonstrated, there is significant variation in speech production ability in children with CIs as they develop speech (Blamey et al., 2001; Ertmer & Goffman, 2011; Flipsen, 2011; Spencer & Guo, 2013). Previous research has identified and classified speech sound errors, created phonetic inventories to illustrate phonological knowledge, and denoted change over time in the accuracy of phoneme production by children with CIs using both broad and narrow transcription (Blamey et al., 2001; Chin, 2003; Flipsen, 2011; Ertmer et al., 2012; Spencer & Guo, 2013). These studies found that, overall, children with CIs develop speech similarly to children with NH. However, some phonemes appear to develop in non-typical ways, and there is no clear explanation for this finding. Examination of covert contrasts in speech sounds produced by both children with CIs and children with NH can shed light on this issue and may have important clinical implications. It is possible that using broad transcription, coupled with a measurement tool that is sensitive to subtle changes in phoneme productions, would demonstrate covert contrast in young children. Since /t/ has been repeatedly shown to be unusually late-developing in children with CIs compared to children with NH (Blamey et al., 2001; Ertmer et al., 2012; Spencer & Guo, 2013), it was

chosen as the phoneme of interest in this investigation. With these factors in mind, there are two research questions that will be addressed:

> 1. Do children with CIs produce /t/ as accurately as children with NH who have similar gross articulatory ability?
> 2. When children with CIs and NH substitute [t] for another sound, are there significant perceptible differences (or covert contrasts) between the /t/ used as a substitution for /d/ or /tʃ/ and typical /t/ productions?

2 Methods

2.1 Speakers

Two groups of preschool-aged children participated in this study: children who used CIs (Experimental Group) and speech-age matched peers (Control Group). All of the children were recruited as part of a larger study that examined the influence of speech production abilities on the speech perception scores of children with CIs (Gonzalez, 2013). Parents of the participants provided the original investigators with detailed demographic information via questionnaire, which allowed them to rule out several exclusionary characteristics. These included: cognitive delay or impairment, cognitive or psychiatric disabilities, and primary language use other than English.

2.1.1 Experimental group

The experimental group included nine congenitally deafened children with profound sensorineural hearing loss (5 females, 4 males) who had been fitted with CIs. All participants in the CI group: 1) were implanted by 3 years of age, 2) had at least 12 months of CI device experience at the time of testing, and 3) used an oral mode of communication exclusively prior to implantation. This was important because previous research has shown that children trained in oral communication have superior consonant acquisition when compared to children with CIs trained with other modes of communication (Connor et al., 2000). Table 1 lists the demographic characteristics of this group.

Table 1. Demographic Characteristics of CI Participants
*1 = High school diploma, 2 = Bachelor's Degree, 3 = Master's Degree/Graduate Certificate, 4 = Doctorate Degree, 999 = did not report
H = Hispanic, C = Caucasian, AA = African American.

ID	Age	Gender	Race/ Ethnic Group	Parent Education**		Age at Implantation (mo)	Age at Activation (mo)	Device Experience (mo)
				Mom	Dad			
CI01	70 mo.	F	H	3	4	21	22	48
CI02	65 mo.	M	C	4	4	8	9	55
CI03	56 mo.	F	C	4	2	14	15	40
CI04	43 mo.	F	AA	2	2	24	26	16
CI05	42 mo.	M	C	3	4	18	19	22
CI06	76 mo.	F	C	2	1	21	21	55
CI07	70 mo.	M	C	2	1	18	20	50
CI08	35 mo.	M	H	3	4	7	8	26
CI09	59 mo.	M	AA	1	999	29	30	28

2.1.2 Control group

Members of the control group were selected from a pool of 24 possible participants. Inclusion criteria were as follows: 1) between the ages of 3-5 years, 2) normal hearing (i.e., hearing thresholds ≤ 20 dB HL from 250 Hz to 4000 Hz), and 3) no middle ear involvement at the time of testing. Of the 24 children whose parents had consented for their child to participate in this study, eight were determined to have appropriate speech production abilities to serve as matches to the experimental group. The control group participants (5 females, 3 males) were between the ages 2:8 to 5:1 years (M = 4:0).

Each child with a CI was matched to a child with NH by articulation ability using scores from a standardized test of articulation and gender, when possible. Raw scores for each participant (i.e., the sum of all articulation errors) were converted into a standard score based on hearing age for the experimental group and chronological age for the control group. Hearing age was defined as time since device activation. Participants were considered "matched" if their respective standard scores fell within the 95% confidence interval of a child with NH (see Table 2). For the NH group, standard score conversions were based on chronological age. Standard scores for the CI group, however, were calculated using the subjects' "hearing age." One matched pair (CI06 and NH17) did not meet this criterion. The standard score for the child with a CI was higher than the NH child based on hearing age, and their 95% confidence intervals did not overlap. However, the two children were exactly the same age (56 months), were both female, and achieved similar raw scores. Given these circumstances, they were considered to have similar articulation abilities and were paired.

Table 2. Matching Criteria for the Participants

Participants with Cochlear Implants							Articulation-Matched, Normal Hearing Participants				
Pairs	ID	Gender	Chron. Age (mo)	Hearing Age (mo)	GFTA-2 SS	GFTA-2 95%CI	ID	Gender	Chron. Age (mo)	GFTA-2 SS	GFTA-2 95%CI
1	CI01	F	70	48	112	106-110	NH15	F	52	108	102-114
2	CI02	M	65	55	103	94-108	NH24	M	49	105	99-111
3	CI03	F	56	40	123	116-130	NH17	F	56	110	104-116
4	CI04	F	43	16	103	97-109	NH11	F	42	105	98-112
5	CI05	M	42	22	121	114-128	NH16	M	43	115	109-121
6	CI06	F	76	55	111	105-117	NH02	F	61	106	101-115
7	CI08	M	35	26	94	87-101	NH23	M	48	100	94-107
8	CI09	M	59	28	103	96-110	NH20	F	32	107	101-113

2.1.3 Listeners

Thirty-three graduate students in speech-language pathology were recruited to participate as listeners in this project. They had completed a phonetics course, voluntarily participated in the listening experiment, and received no compensation.

2.2 Materials

Speech and language data were obtained using the Peabody Picture Vocabulary Test 4 (PPVT-4; Dunn & Dunn, 2007), the Goldman-Fristoe Test of Articulation-2 (GFTA-2; Goldman & Fristoe, 2000), and the On-line Imitative Test of Speech-Pattern Contrast Perception (OlimSpac; Boothroyd et al., 2010). The first two measures reflected speech and language ability. Speech samples were taken from the OlimSpac. This computerized software program provides a measure of speech perception by assessing the production of six phonologically significant speech contrasts in children with hearing loss (see Table 3).

During OlimSpac testing, pre-recorded VCV nonwords were presented over a loudspeaker, while the child was seated in front of a computer monitor in a sound-proof booth. The child was instructed to "watch the screen", listen for each sound presentation, and repeat the nonsense word to the best of their ability. Each OlimSpac stimulus item was presented to the child in both an auditory-only and auditory-visual condition. During the auditory-only trials, the screen displayed a colorful image that changed color when the stimulus played. During the auditory-visual trials, the screen displayed an adult female's face as she pronounced the stimulus accurately. Each speech contrast was represented at least twice by different phonemes. Selected contrasts were consistent among subjects but presented in a random order during each test session. Each child imitated 16 VCV nonwords in each condition (auditory-visual, and auditory-only), for a total of 32 imitated productions per child. The children's imitated productions were recorded for future analysis using an Olympus ME52

directional lapel microphone connected to an RCA VR 5220 digital voice recorder. These productions served as the acoustic stimuli for the current investigation.

Table 3. OlimSpac Speech Contrasts

Speech Contrast	Example
Vowel height	/udu/ vs. /ada/
Vowel place	/utu/ vs. /iti/
Consonant voicing	/ata/ vs. /ada/
Consonant continuance	/iti/ vs. /isi/
Pre-alveolar consonant place	/upu/ vs. /utu/
Post-alveolar consonant place	/utu/ vs. /uʧu/

2.2.1 Development of experimental protocol

For this project, a graduate student in speech-language pathology (SLP) phonetically transcribed subject responses, from the GFTA-2 which were then reviewed by a second graduate SLP student. A third "expert" clinician, who was a certified SLP, was consulted to resolve discrepant transcriptions and made the final decision. These transcriptions and OlimSpac recordings then were analyzed by the second author, who did not participate in the testing of the participants or scoring of the GFTA-2. She determined whether the VCV syllables represented a correct production or a clear substitution. Distortions were counted as correct, despite mild phonetic differences (inappropriate aspiration, imprecise production, etc).

The investigators selected /t/, /d/, and /ʧ/ as the phoneme productions of interest. These phoneme choices were particularly appropriate because one differed in voicing ([t] for /d/) and the other in manner of articulation ([t] for /ʧ/). Place of articulation consistently has been shown to be poorly transmitted by CIs (Clark, 2003; Giezen et al., 2010; Pisoni et al., 1999), so a place contrast (such as [t] for /p/) was not included in this experiment. In addition, since coronal place of articulation has been shown to be well-transmitted by the speech processors of CIs, one can assume that the speakers in this study received as much acoustic information as possible from their speech processors for adequate /t, d, ʧ/ perception (Dillon et al., 2004).

For each subject, every opportunity for the three target consonants was isolated and digitized at 20,000 Hz using Praat (Boersma & Weenink, 2013). Each child had eight opportunities to produce /t/, and four opportunities each to produce both /d/ and /ʧ/. The following VCV contexts were utilized: /ata/, /utu/, /iti/, /ada/, /udu/, /iʧi/, and /uʧu/. No effort was made to control for listening condition because the original investigators found no significant difference in consonant accuracy between the auditory-only and the auditory-visual conditions for either the NH or CI group.

The selected files underwent noise reduction using Audacity® (SourceForge, 2013) and were then normalized.

The prepared sound files were divided into two blocks: all samples of target [d] were placed into block 1, and all samples of target [ʧ] were placed into block 2. Samples of target [t] were equally distributed between the two blocks. Each block contained 60 unique speech production samples evenly distributed across CI and NH children. File order was quasi-randomized to ensure that no more than two similar-sounding files, (either by stimulus or subject) were presented consecutively. The first 12 files presented in each block were duplicated for presentation at the end of the block in order to assess intra-rater reliability.

Although previous studies on covert contrasts had used visual analogue scales (VAS), this experiment used equal-appearing interval scales (EAI scales). According to Yiu and Ng (2004), EAI scales showed significantly higher intra-rater reliability than VAS (EAI agreement = 0.73; VAS agreement = 0.57), and there was a moderate correlation (.56-.76) between EAI and VAS scale ratings for identical stimuli. Since consistent judgments are essential when assessing a child's progress toward a target sound, the use of an EAI scale should produce similar results to VAS and was used in this experiment.

2.3 Procedures

When the listeners arrived to participate in the study, they were asked to fill out a brief questionnaire in order to ensure consistency in listener characteristics. All listeners self-reported: adequate hearing, typical neurological status and cognition, and English as a first language. Additionally, no listener showed evidence of a speech or language disorder, as judged by the examiner.

ECoS Win experimental design software (Avaaz, 2002) was used to present the experimental trials on a Dell Optiplex desktop using Califone circumaural headphones. Each experimental block was preceded by a training block consisting of 10 novel sound files that were not utilized in the experimental blocks. The listeners were told that they would be listening to children producing VCV nonwords and were given an example (like [ada]). Then, listeners were shown a 7-point EAI scale. They were asked to click a point on the scale that most closely corresponded to their interpretation of the phonetic accuracy of the consonant presented in each trial. A score close to either extreme of the EAI indicated a very accurate production of a phone, with 1 or 7 being a "perfect" production of that phone. A score of 4 would represent an inability to distinguish between the two phonemes. In block 1, listeners rated the subjects' attempts at /t/ and /d/. In block 2, they rated attempts at producing /t/ and /ʧ/.

3 Results

3.1 Intra-rater reliability

Over both experimental blocks, each listener rated 12 stimuli twice (N = 48 trials). For each listener, the percentage of responses to duplicated stimuli that

were within ± 1 scale value of the original rating was calculated (Kreiman et al., 1993). These values were averaged across listeners. Calculations revealed that overall, 88.1% of duplicated trials were within ± 1 scale value of the original rating. Of these, 56.6% were in exact agreement. Hence, listener reliability was determined to be very good.

3.2 Statistical analyses

A three-way repeated measures ANOVA was conducted to analyze the influence of group (CI vs NH), Transcription Category (4 levels of correct and substituted productions), and Covert Contrast Category (/d/ or /ʧ/) on perceptual ratings. Results revealed a significant three-way interaction. However, differences across the experimental blocks were not of primary interest and will not be discussed further. Statistical analysis also revealed that two of the three main effects were significant, experimental group, $F(1,32) = 27.99$, $p < .001$, $\eta_p^2 = 0.467$ and transcription category (TC), $F(3,96) = 760.70$, $p < .001$ $\eta_p^2 = 0.960$. These results suggest that differences were found across both groups and TC. However, the statistically significant interaction between group and TC was of particular importance, therefore, the research questions will be addressed within this interaction.

3.2.1 Accuracy of /t/ productions

In the past, /t/ has been shown to be unusually late developing in children with CIs (Blamey et al., 2001; Chin, 2003; Ertmer et al., 2012). The first goal of this project was to confirm this observation by examining listener perceptions of [t] accuracy. This was best satisfied by examination of the significant Group x TC interaction, $F(3,96) = 25.562$, $p < .001$, $\eta_p^2 = 0.444$. This finding suggests that differences in transcription category were dependent upon group. Post-hoc testing results using paired samples t-tests with a Bonferroni correction ($p = .004$) revealed that 3 out of 8 paired comparisons of interest were not significant: [t] for /t/ in both experimental blocks, and [ʧ] for /ʧ/ (see Figures 1 and 2). In other words, [t] and [ʧ] productions were similarly accurate across groups; however [d] for /d/ productions were significantly more accurate in children with NH. Hence, when the production was judged to be a /t/ by SLPs, children with CIs successfully produced /t/ as accurately as their NH peers.

While the above findings demonstrated no group differences for /t/, it did not address the issue of whether or not /t/ was produced in error by children with CI more often than other phonemes or when compared to [t] productions from NH children. To test this hypothesis, overall error frequency taken from the OlimSpac testing was determined to provide enough additional relevant information to warrant analysis. A confusion matrix of CI group productions had previously been generated when selecting contrastive consonant choices. To

compare error frequencies between groups, a second confusion matrix (of NH productions) was created (see Table 4).

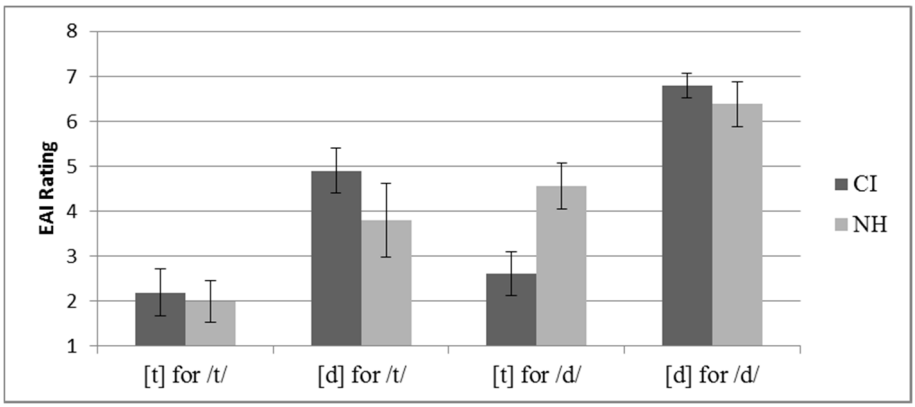


Figure 1.
Differences in listener perceptions of consonant accuracy
for [t] for /t/ and [d] for /d/.
*Covert contrast is shown in the [d] for /t/ and [t] for /d/ contrasts.

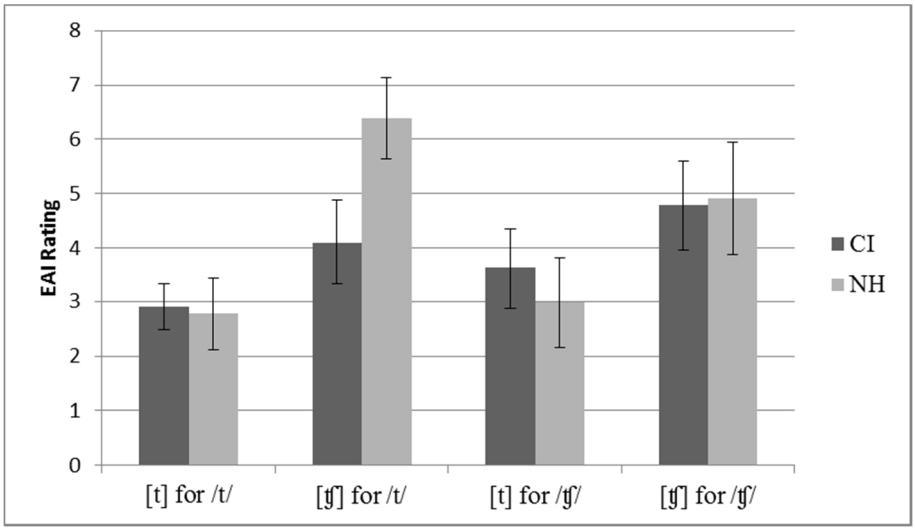


Figure 2.
Differences in listener perceptions of consonant accuracy
for [t] for /t/ and [ʧ] for /ʧ/.
*Covert contrast is shown in the [ʧ] for /t/ and [t] for /ʧ/ contrasts.

Table 4. Confusion matrix for responses on the OlimSpac produced by children with CI vs. NH

OlimSpac Target	Child's Production CI(top) NH(bottom) /p/	/b/	/d/	/t/	/s/	/ʃ/	/ʧ/	Other
/p/	72.22 81.25	2.78 15.63	8.33	8.33				8.33 3.13
/b/	13.89 9.38	63.89 78.13	11.11 6.25	8.33 3.13				2.78 3.13
/d/		 3.13	62.16 81.25	27.03 12.50				10.81 3.13
/t/		1.41	7.04 8.62	73.24 87.93	1.41	2.82	14.08 3.45	
/s/	2.78		2.78	2.78	52.78 68.75	22.22 9.38		16.67 21.88
/ʃ/			5.56	5.56	2.78 21.88	66.67 71.88	13.89 3.13	5.56 3.13
/ʧ/			2.78	19.44 21.88		22.22 6.25	55.56 71.88	

Examination of this confusion matrix revealed that children with NH produced /t/ accurately in 87.93% of opportunities, whereas children with CIs produced it accurately in 73.24% of opportunities. Hence, [t] was found to be perceptually less accurate in children with CI when compared to those [t]s produced by children with NH. Nevertheless, [t] was the most accurate phoneme produced by the children with CI when compared to the other OlimSpac test stimuli, which also matched the performance of the NH group.

3.2.2 Perceptible contrasts between substitutions and correct targets

The second purpose of this investigation was to determine whether or not covert contrast was present in the speech of children with CIs, and if so, were the patterns of covert contrast similar to those observed in children with NH? Identification of covert contrast was best addressed by an examination of the significant within group post hoc results of the Group x TC interaction. All within group paired comparisons for both the CI and NH group were significant. In other words, the "correct" /t/ was rated significantly different from the [t] used as a substitution, as well as contrasting with [d] and [ʧ] when they were used as a substitute for a /t/. In addition, post hoc testing revealed significant differences in the similarity of [t, d, ʧ] productions across groups when they were used as substitutions for other phonemes (i.e., [t] for /d/, [d] for /t/, [t] for /ʧ/, [ʧ] for /t/). This finding suggests that there were significant differences in the patterns of covert contrast across groups. As illustrated in Figures 1 and 2, all

four paired comparisons involving phoneme substitutions across groups were significant ($p < .001$). When children with CIs substituted [d] for /t/, it was perceived as more [d]-like and when they substituted [t] for /d/, it was perceived as more [t]-like. The opposite pattern was noted in NH children. However, a different tendency was noted for /ʧ/. For children with CIs, the [t] for /ʧ/ substitution was more /ʧ/-like than for NH hearing children. While there was a significant group difference for the [ʧ] for /t/ substitution, there was only one instance of this error in the NH group, so a group comparison is not appropriate. Nevertheless, the [t] production in this condition for the children with CIs was more [ʧ]-like.

4 Discussion

The current results suggest the measurement technique used by the listener does influence the reporting of developmental speech patterns for children with CIs. When using phonetic transcription, the children with CIs were less accurate in phoneme production than speech age-matched children with NH. However, when EAI scales were used to rate the same speech productions, listeners identified different patterns of covert contrast across these groups.

4.1 The development of /t/ in children with CIs

The first research question dealt with the accuracy of /t/ production when the listener decision of phonetic accuracy of /t/ across speaker groups (CI vs. NH) varied by technique: EAI scale versus phonetic transcription. A three-way repeated measures ANOVA using the data from the EAI scale revealed that listeners perceived no significant difference between groups when only the correct /t/ productions were considered. Hence, children using CIs were no less accurate in their [t] productions than children with NH when speakers were matched for articulation ability.

These non-significant findings are likely related to advances in CI speech processing technology, as the previous studies that showed delayed /t/ acquisition were conducted over 10 years ago (Blamey et al., 2001; Chin, 2003). Another possible explanation involves device experience. The two previous studies that revealed delayed /t/ development tested participants with less than 3 years of device experience (Ertmer et al., 2012; Spencer & Guo, 2013). The children who participated in this project averaged three years of device experience. Given that children with CIs acquire speech sound accuracy quickly at first, and then slow down, it is possible that our participants were in the "plateau" stage, given their length of device experience and history of oral language use, whereas those in the comparison studies were still in the early stages of development, characterized by rapid growth in their phonetic inventories.

The second analysis of the accuracy of /t/ was derived from an examination of all phonemes tested on the OlimSpac. Error proportions for each participant were collapsed by group and placed in a confusion matrix. Results indicated that children with CIs made more speech sound errors than children with NH for all phones tested, including /t/. However, /t/ was not significantly more impaired than the other phonemes produced by children with CI. Interestingly, both groups produced /t/ accurately more often than other phonemes evaluated on the OlimSpac (e.g., /p, b, d, s, ʃ, ʧ/). These findings do not support those of Ertmer et al. (2012) who reported that initial /t/ was less accurate than both /d/ and /ʧ/ in children with CIs during acquisition.

4.2 Use of covert contrast

The second research question addressed the presence of covert contrast in children with CIs. To address this issue, listener ratings of /t, d, ʧ/ substitutions were compared with ratings of correct tokens for the same phonemes. Covert contrast was present if the two sounds (one substituted, one correct) were transcribed identically but rated differently by listeners on the EAI scale. If covert contrast was present, then the child was in the process of developing the speech sound. If not, then the error suggested lack of phonological knowledge for the target phoneme contrast.

Post-hoc comparisons of the group by transcription category (TC) interaction showed that both CI and NH groups produced perceptible differences between correct productions and substitutions for the target phonemes (Figures 1 and 2). The voicing contrast was more readily perceived in the productions by children with NH while the children with CIs struggled with this contrast. That is, the [t] for /d/ substitutions produced by children with CIs sounded more like [t] and [d] for /t/ substitutions sounded like [d]. These errors lack covert contrast and support difficulties with voicing. This finding confirms Gonzalez's (2013) conclusion that the children with CIs struggled with the perception and/or production of voicing more than with other phoneme distinctions. Since the OlimSpac uses VCV syllables, one might expect more difficulties with syllables that alternate voicing (/ata/) than one that is entirely voiced (/ada/). It was surprising that children with CIs struggled with both syllable types.

A comparison could not be made across groups for the manner (plosive/affricate) contrast since so few [ʧ] for /t/ errors were noted in the NH group. However, the children in the CI group produced a sufficient number of both [t] for /ʧ/ and [ʧ] for /t/ errors for analysis. Results indicated that those with CIs had good phonological representations for /t/, and the production of covert contrast in the errors revealed productions closer to the desired target, either [t] or [ʧ]. This demonstration of covert contrast in children with CIs supports the idea that they have acquired both [t] and [ʧ] but have not completely mastered either.

An interesting finding was that the children with CIs were able to produce a perceptible contrast between correct [t] and the [t] for /ʧ/ substitutions, while children with NH did not. As expected for children with NH who were approximately 4;0 years old, they did not have a mature phonological representtation for /ʧ/, as it is a later developing phoneme. On the other hand, unlike their NH peers, the children in the CI group, with an average chronological age of 4;9 years, were developing this contrast. Even though the CI group only had an average of 3;2 years of robust hearing experience, they were at the approximate chronological age for the development of /ʧ/ (Smit et al., 1990). Since these results are based on listener perceptions of covert contrast, it is possible that children with CIs use a certain speech feature (like aspiration or voice onset time) to make [t] substitutions sound more like [t] when contrasted with /d/, and more like [ʧ] when attempting to produce /ʧ/. Hence, the use of a CI might influence which acoustic cues the child attends to in the development of phonemic contrasts, or it is possible that these children weigh the available cues in a different way than children with NH do. More detailed acoustic analyses are needed to test these hypotheses.

4.3 Clinical implications

This investigation has shown that subtle differences in phoneme accuracy are often perceptible by an experienced listener. A clinician who is able to reliably gauge the presence and extent of covert contrast may be able to provide more accurate prognostic statements and select treatment targets that will facilitate student progress.

There are two different ways to select a target for children with NH (Gierut, 2007; Miccio, 2005). Based on the child's learning style, the clinician can choose a target for which the child has contrastive knowledge (i.e., a sound produced with covert contrast) or one that is unknown to the child. In other words, is phonetic accuracy or the learning of a new phonemic contrast the focus of treatment? Since research has demonstrated the utility of narrow transcription in the identification of speech sound errors in children with CIs (Teoh & Chin, 2009), it may be possible to incorporate an assessment of covert contrast in an evaluation of speech sound disorder so that treatment decisions can be enhanced. The current study indicates that covert contrast can provide the data necessary to make decisions about target selection and that covert contrast can be used to track progression towards phoneme mastery.

Acknowledgments

This paper is based on the Master's thesis of the second author.

References

Avaaz Innovations Inc. (2002). *ECoS Experiment Generator 2.0* [computer software].

Blamey, P. J., Barry, J. G., & Pascale, J. (2001). Phonetic inventory development in young cochlear implant users 6 years postoperation. *Journal of Speech, Language, and Hearing Research, 44,* 73-79.

Boersma, P., & Weenink, D. (2013). *Praat: Doing phonetics by computer* (version 5.3.56). The Netherlands: University of Amsterdam. Retrieved from www.praat.org October 2013.

Boothroyd, A., Eisenberg, L. S., & Martinez, A. S. (2010). An on-line imitative test of speech-pattern contrast perception OlimSpac: Developmental effects in normally hearing children. *Journal of Speech, Language, and Hearing Research, 53,* 531-542.

Byun, T. M., Buchwald, A., & Mizoguchi, A. (2016). Covert contrast in velar fronting: An acoustic and ultrasound study. *Clinical Linguistics & Phonetics, 30,* 249-276.

Connor, C. M., Hieber, S., Arts, H. A., & Zwolan, T. A. (2000). Speech, vocabulary, and the education of children using cochlear implants: Oral or total communication? *Journal of Speech, Language, and Hearing Research, 43,* 1185-1204.

Chin, S. B. (2003). Children's consonant inventories after extended cochlear implant use. *Journal of Speech, Language, and Hearing Research, 46,* 849-862.

Clark, G. (2003). *Cochlear Implants: Fundamentals and Applications.* New York: Springer.

Dillon, C. M., Cleary, M., Pisoni, D. B., & Carter, A. K. (2004). Imitation of nonwords by hearing-impaired children with cochlear implants: Segmental analyses. *Clinical Linguistics and Phonetics, 18,* 39-55.

Dunn, D. M., & Dunn, L. M. (2007). *Peabody Picture Vocabulary Test, Fourth Edition.* Minneapolis: NCS Pearson, Inc.

Ertmer, D. J., & Goffman, L. A. (2011). Speech production accuracy and variability in young cochlear implant recipients: Comparisons with typically developing age-peers. *Journal of Speech, Language, and Hearing Research, 54,* 177-189.

Ertmer, D. J., Kloiber, D. T., Jung, J., Kirleis, K. C., & Bradford, D. (2012). Consonant production accuracy in young cochlear implant recipients: Developmental sound classes and word position effects. *American Journal of Speech-Language Pathology, 21,* 342-353.

Flipsen, P. (2011). Examining speech sound acquisition for children with cochlear implants using the GFTA-2. *The Volta Review, 11,* 25-37.

Forrest, K., Weismer, G., Hodge, M., Dinnsen, D. A., & Elbert, M. (1990). Statistical analysis of word-initial /k/ and /t/ produced by normal and phonologically disordered children. *Clinical Linguistics & Phonetics, 4,* 327-340.

Gierut, J. (2007). Phonological complexity and language learnability. *American Journal of Speech-Language Pathology, 16,* 6-17.

Giezen, M. R., Escudero, P., & Baker, A. (2010). Use of acoustic cues by children with cochlear implants. *Journal of Speech, Language, and Hearing Research, 53,* 1440-1457.

Goldman, R., & Fristoe, M. (2000). *Goldman Fristoe Test of Articulation—Second Edition.* Minneapolis, MN: Pearson Assessments.

Gonzalez, V. (2013). *Effects of Speech Production Ability on a Measure of Speech Perception Capacity in Young Children with Cochlear Implants and their Articulation-Matched Peers*. Unpublished dissertation completed at the University of South Florida.

Hewlett, N., & Waters, D. (2004). Gradient change in the acquisition of phonology. *Clinical Linguistics and Phonetics, 18,* 523-533.

James, D., Brinton, J., Rajput, K., & Goswami, U. (2008). Phonological awareness, vocabulary, and word reading in children who use cochlear implants: Does age of implantation explain individual variability in performance outcomes and growth? *Journal of Deaf Studies and Deaf Education, 13,* 117-137.

Kirby, J. (2011). Modeling the acquisition of covert contrast. In *Proceedings of the 17th International Congress of Phonetic Sciences* (pp. 1090-1093). Hong Kong.

Kreiman, J., Gerratt, B. R., Kempster, G. B., Erman, A., & Berke, G. S. (1993). Perceptual evaluation of voice quality: Review, tutorial, and a framework for future research. *Journal of Speech, Language, and Hearing Research, 36,* 21-40.

Li, F., Edwards, J., & Beckman, M. E. (2009). Contrast and covert contrast: The phonetic development of voiceless sibilant fricatives in English and Japanese toddlers. *Journal of Phonetics, 37,* 111-124.

Macken, M. A., & Barton, D. (1980). The acquisition of the voicing contrast in English: A study of voice onset time in word-initial stop consonants. *Journal of Child Language, 7,* 41-74.

Miccio, A. (2005). A treatment program for enhancing stimulability. In A. G. Kamhi, & K. E. Pollock (Eds.), *Phonological Disorders in Children* (pp. 163-173). Baltimore, MD: Paul Brookes Publishing Co.

Munson, B., Edwards, J., Schellinger, S. K., Beckman, M. E., & Meyer, M. K. (2010). Deconstructing phonetic transcription: Covert contrast, perceptual bias, and an extraterrestrial view of Vox Humana. *Clinical Linguistics & Phonetics, 24,* 245-260.

Munson, B., Schellinger, S. K., & Carlson, K. U. (2012). Measuring speech-sound learning using visual analog scaling. *SIG 1 Perspectives on Language Learning and Education, 191,* 19-30.

Munson, B., Schellinger, S. K., & Edwards, J. (2017). Bias in the perception of phonetic detail in children's speech: A comparison of categorical and continuous rating scales. *Clinical Linguistics & Phonetics, 31,* 56-79.

Pisoni, D. B., Cleary, M., Geers, A. E., & Tobey, E. A. (1999). Individual differences in effectiveness of cochlear implants in children who are prelingually deaf: New process measures of performance. *The Volta Review, 101,* 111-164.

Schellinger, S. K., Munson, B., & Edwards, J. (2017). Gradient perception of children's productions of /s/ and /θ/: A comparative study of rating methods. *Clinical Linguistics & Phonetics, 31,* 80-103.

Scobbie, J. M., Gibbon, F., Hardcastle, W. J., & Fletcher, P. (1996). Covert contrast as a stage in the acquisition of phonetics and phonology. *QMC Working Papers in Speech and Language Sciences, 1,* 43-62.

Serry, T. A., & Blamey, P. J. (1999). A 4-year investigation into phonetic inventory development in young cochlear implant users. *Journal of Speech, Language, and Hearing Research, 42,* 141-154.

Smit, A. B., Hand, L., Freilinger, J. J., Bernthal, J. E., & Bird, A. (1990). The Iowa articulation norms project and its Nebraska replication. *Journal of Speech and Hearing Disorders, 55,* 779-798.

SourceForge (2013). *Audacity* (Version 2.0.5) [Computer software].Retrieved from audacity. sourceforge.net/download October, 2013

Spencer, P. E. (2002). Language development of children with cochlear implants. In P. E. Spencer (Ed.), *Cochlear Implants in Children: Ethics and Choices* (pp. 222-249). Washington, D. C.: Gallaudet University Press.

Spencer, L. J., & Guo, L. Y. (2013). Consonant development in pediatric cochlear implant users who were implanted before 30 months of age. *Journal of Deaf Studies and Deaf Education, 18,* 93-109.

Strombergsson, S., Salvi, G., & House, D. (2015). Acoustic and perceptual evaluation of category goodness of /t/ and /k/ in typical and misarticulated children's speech. *Journal of the Acoustical Society of America, 137,* 3422-3435.

Teoh, A., & Chin, S. (2009). Transcribing the speech of children with cochlear implants: Clinical application of narrow phonetic transcription. *American Journal of Speech Language Pathology, 18,* 388-401.

Tyler, A. A., Figurski, G. R., & Langsdale, T. (1993). Relationships between acoustically determined knowledge of stop place and voicing contrasts and phonological treatment progress. *Journal of Speech and Hearing Research, 36,* 146-759.

Yiu, E., & Ng, C. (2004). Equal appearing interval and visual analogue scaling of perceptual roughness and breathiness. *Clinical Linguistics & Phonetics, 18,* 211-229.

REDEFINING CONCATENATIVE SPEECH SYNTHESIS FOR USE IN SPONTANEOUS CONVERSATIONAL DIALOGUES; A STUDY WITH THE GBO CORPUS

Nick CAMPBELL
Speech Communication Lab, School of Computer Science & Statistics,
Faculty of Mathematics & Engineering, Trinity College Dublin,
The University of Dublin, Ireland
nick.campbell@tcd.ie

Abstract
This chapter describes how a very large corpus of conversational speech is being tested as a source of units for concatenative speech synthesis. It shows that the challenge no longer lies in phone-sized unit selection, but in categorising larger units for their affective and pragmatic effect. The work is by nature exploratory, but much progress has been achieved and we now have the beginnings of an understanding of the types of grammar and the ontology of vocal productions that will be required for the interactive synthesis of conversational speech. The chapter describes the processes involved and explains some of the features selected for optimal expressive speech rendering.

Keywords: unit selection, conversational speech, feature categories, corpus processing, spontaneous interaction

1 Introduction

Speech Synthesis has moved from being a research issue to a service that is being provided by businesses for businesses worldwide (Capes et al., 2017; Wan et al., 2017; Pollet et al., 2017). There are still many active research topics remaining, but the technology can now be considered mature. For most business applications, a consistent voice is the main requirement; i.e., one that can be 'branded' to convey the desired 'company image' for e.g., Call Centre applications or Customer Care services. For Assisted Living, on the other hand, it might be more appropriate to employ a voice that changes its quality with different situations, sounding 'strict' at some times but 'soft and caring' at others (Sorin et al, 2017; Gilmartin et al, 2018).

No single voice can in practice be good at everything; a news-reader voice might not be optimal for poetry reading for example, but 'expressivity' has become a major research area for synthetic voice creation (Campbell, 2004;

DOI: http://doi.org/10.18135/CAPSS.157

Abadi et al., 2016; Wang et al., 2016; Arik et al., 2017). For this, a representative corpus that illustrates the scope of vocal variation in everyday interactive situations is essential.

The following sections will describe one such corpus, and a synthesis system capable of using it, and will outline the steps and challenges of the work. We present a unique 600-hour corpus of one speaker recorded systematically over a period of 5 years, throughout which she encountered many and various interlocutors and situations, resulting in a database of recordings that might eventually become the world's largest synthetic voice. But first, we must develop a science of situated speaking styles that accounts for the vocal variation it illustrates, and an ontology of speech sounds that are frequent and ubiquitous but that never occur as entries in any language dictionary.

2 The GBO Corpus

The GBO Corpus (Guttural Behaviour Ontology) is a set of recordings made over a period of five years as part of the JST Expressive Speech Processing project (Campbell, 2001). The data were never released because of personal privacy considerations, although full legal rights to use the material and to make it public for research purposes were granted freely and with informed consent by the speaker both at the onset of the recordings and after their completion. The name comes from the remarkable finding that almost half of the speech sounds were 'non-lexical' or 'guttural' noises that function as normal sounds in casual spoken interaction but that are not typically found in a dictionary of the formal language. These sounds form perhaps the biggest challenge to 'conversational' or 'interactive' speech synthesis as they are so hard to specify in text, and so easy to misinterpret if badly or inappropriately generated.

The recordings were made using a professional-quality head-mounted microphone and stored to MiniDisk. They were purchased by the project from the speaker on a regular monthly basis as part of the ESP corpus collection between the years

2000 and 2005 were inclusive. There were many speakers employed over this period, but GBO (name concealed) was remarkable in the quantity and quality of her recordings. The speaker had full rights to withhold any material and was of course able to monitor the content and self-censor before bringing her data to the lab for our research. Nonetheless, the recordings contain some very personal information and after they were individually manually transcribed by third-party specialists (part of our team who had signed confidentiality agreements), we decided on moral grounds that they should not be made public, out of respect for the speaker and her personal privacy. GDPR may now facilitate their research use under strict confidentiality.

However, this resource yields priceless information for the generation of conversational speech synthesis for an interactive spoken dialogue system, such as might be used in assistive living or customer-care applications. Because the speaker recorded virtually everything she said (in exchange for an income well above the minimum wage while doing so) we have a unique sample of the everyday speech of one person in a variety of daily-life interactions over an extended period of time.

3 CHATR High Definition Speech Synthesis

The CHATR speech synthesis system was developed throughout the early nineties in Kyoto, Japan, in Department 2 of the now defunct ATR Interpreting Telephony Labs (later Interpreting Telecommunications Research Labs) and was announced in 1996 as "a high-definition speech re-sequencing system" at the joint ASJ/ASA meeting in Hawaii (Campbell, 1996) and at ICASSP in the same year (Hunt & Black, 1996), though the basic method was first reported in 1994 at the ESCA/IEEE Mohonk Speech Synthesis workshop (Campbell, 1994). The name was derived from Collected Hacks from ATR and was first suggested by Paul Taylor who was then working on the intonation component. It was not the first concatenative speech synthesis system (see e.g., Moulines & Charpentier, 1990; Sagisaka et al., 1992) but it was the first to use raw waveform segments directly, without recourse to any signal processing. This step not only greatly simplified the synthesis process but also allowed the use of very high quality recordings (some even in stereo) that exactly reproduced the voice quality and speaking style of the recorded subjects. It replaced the buzzy artificial sound of parametric synthesis with surprisingly natural-sounding speech. It was susceptible to concatenation errors if the waveform coverage in the voice database was incomplete but in that period much progress was made using as little as one hour of recorded speech and the samples in the corpus are all produced from such small databases. In contrast, some commercial users of this system now employ corpora of well-over 100 hours of recordings.

4 CHATR & GBO

This section reports ongoing work to synthesise conversational speech from the GBO corpus using CHATR technology. It describes the steps that are required to reduce the candidate segments when searching in such a large database, and the features that can be used to maximise expressivity in the speech. The largest databases for unit concatenation synthesis to date have been specially recorded using professional voice talent over an extended period of up to about 150 hours. These pro fessionals are capable of maintaining the same tone of voice throughout all recordings and can provide a large and consistent database of speech samples. Our GBO speaker, on the other hand, was recorded in a range of activi-

ties throughout her daily life and of course made no conscious effort to maintain any consistency in her voice. In fact she changed her speaking style and tone of voice consistently when talking with different people. She was 'not the same person' when talking with her parents as when talking with her bank manager for example. This is precisely the component that we wish to make use of in 'interactive speech synthesis' for 'spontaneous' conversations in interactive dialogues.

The entire corpus was manually transcribed into utterance units, and half the corpus was manually annotated for speaking style and speaker state, in addition to interlocutor information for each utterance. We therefore have an index of suprasegmental information that can be used to influence the selection of segments for concatenation. Figure 1 provides an example of the raw metadata. There are 266,599 manually labelled utterance entries of this sort in the GBO corpus.

```
GBO018_01,9.666,10.587,（親との電話）［X］,ci-rough,0,no,,,,,,,
GBO018_01,21.341,22.077,もしもし,ka-rough,m1,yes,,,,,,,
GBO018_01,22.349,22.735,うん,ka-rough,m1,yes,,,,,,,
GBO018_01,22.947,24.245,絨毯王国,ka-rough,m1,yes,,,,,,,
GBO018_01,25.035,26.678,あの一,ka-rough,m1,yes,,,,,,,
GBO018_01,27.852,29.142,阪奈から行つたら,ka-rough,m1,yes,,,,,,,
GBO018_01,30.906,32.033,あの一,ka-rough,m1,yes,,,,,,,
GBO018_01,32.702,37.01,途中で曲がんねんやん、ジャパンの筋、ジャパンとココスの筋を曲がつて,ka-rough,m1,yes,,,,,,,
GBO018_01,46.011,49.506,あのなあ、（西大寺）［さいだいじ］奈良ファミリーんとこ過ぎて,ka-rough,m1,yes,,,,,,,
GBO018_01,53.396,54.741,うん、ほんて,ka-rough,m1,yes,,,,,,,
```

Figure 1.
Sample of annotations for GBO, showing file-id, start-time, end-time, utterance text, interlocutor-id, voice-quality, manner of speaking, etc., as noted in csv format

4.1 Corpus Processing

The first challenge in processing the entire corpus for 'spontaneous' speech synthesis is to extend the suprasegmental annotation across the whole corpus by training on the manually-produced portion and automatically generating information for the remainder of the unlabelled utterances by statistical processing. Although this is now a standard procedure, several sub-challenges need to be solved before it can be done properly – these include determining the optimal form for the units (their granularity) and the optimal features by which to index them. Figure 2 illustrates the flow of this processing. On the left of the figure we see the preparatory grammar learning processes (a1 a4), and on the right the extended unit-selection database processing. In the middle are the original corpus (a) transcribed in Japanese kana-kanji text, with special diacritics and symbols for vocal productions which are not well characterised in writing (laughs, lip-noises, and expressive interjections for example), and (b) the

resulting multigram (Deligne & Bimbot, 1997) database of optimal symbolic representtations produced from the work.

On the right (b1b5), we see the flow of feature-based indexing by which the unit database is annotated for retrieval of appropriate speech tokens. Acoustic features no longer need to be represented directly in the index, as many of their characteristics are direct consequences of the higher-level constraints such as interlocutor identity and utterance pragmatics, which can be used as selection criteria in the unit selection process (see Section 5.2). The prosodic and voice-quality feature weights are therefore calculated from correlations with the higher-level predictors. These in turn are now required as part of the input for utterance selection.

Table 1. Almost half of the GBO utterances were found to be 'non-lexical', or 'guttural vocalisations' not to be found in a typical dictionary of the spoken language

total number of GBO utterances transcribed:	148,772
number of unique lexical utterances:	75,242
number of non-lexical utterances:	73,480
number of non-lexical utterance types:	4,492
proportion of non-lexical utterances:	49.4%

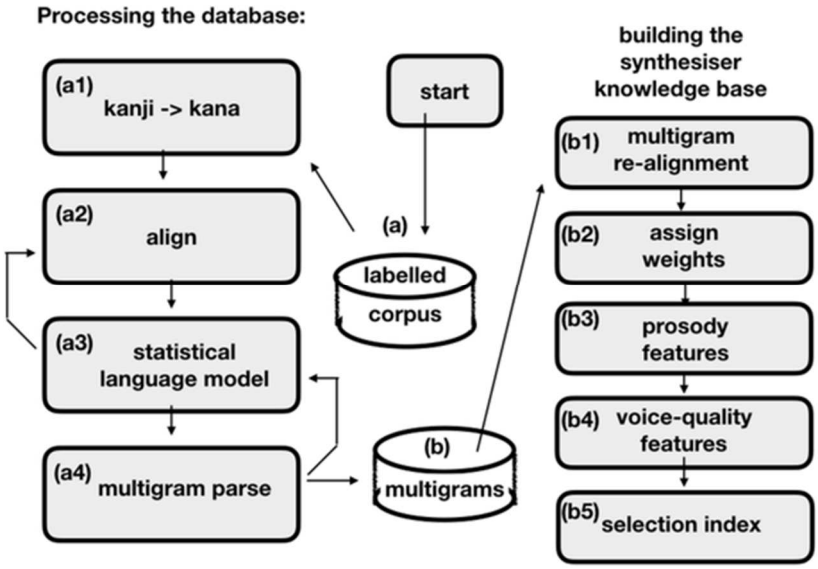


Figure 2.
Flow of processing of the GBO corpus for use in CHATR synthesis

4.2 Synthesis Generation

The original CHATR software featured two research-level modules that were not much spoken of at the time, but have proved remarkably insightful for the

processing of massive conversational data. The UnitMan module was originally designed by Patrick Davin as a debugger to test the weight-based selection of units in the corpus by manually exploring the selection-space and enabling listening to closely aligned candidate segments that emerged through the selection process. PhraseBank was designed originally to manually 'correct' any utterances that were not properly rendered by the default weights in unit selection; i.e., to be able to produce and proactively store utterances that were required as output but known not to be ideal when generated by synthesis automatically. These modules have proven especially useful for the present work.

Figure 3 shows a screen printout when both modules are being used interactively. The utterance 'koNnichiwa' ('Hello' in Japanese) has been selected using phone, syllable, and word-sized units that were in the database, with the final selection marked in red saved as a phrase with a given name. The same text might require several renderings when produced in different situations 'Hello' as a greeting, as a call, as an exclamation, or as a citation, for example, and these different renderings can be given unique IDs to be used as pre-stored phrases for quick generation of the appropriate sound.

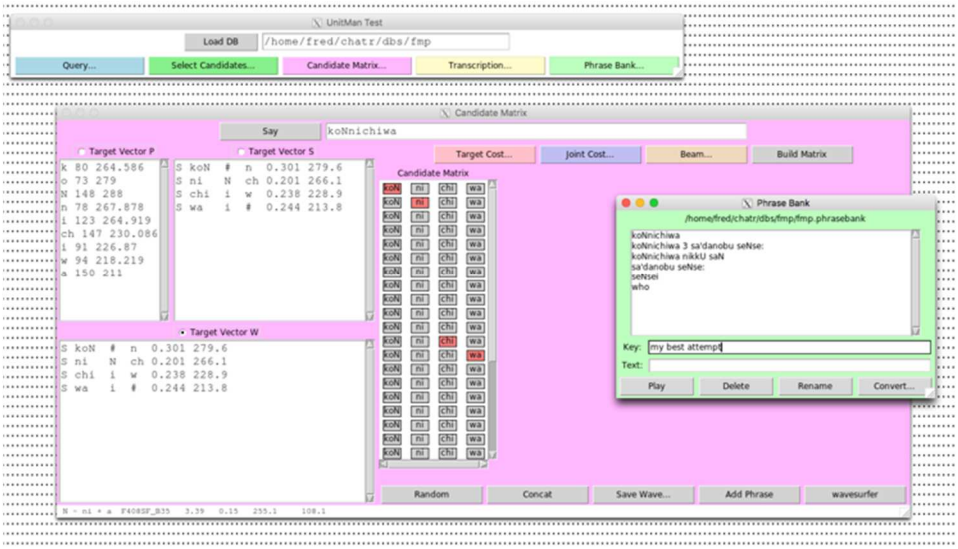


Figure 3.
CHATR's UnitMan and PhraseBank modules

5 Optimal Units for Synthesis

When the source-speech data were still relatively sparse, 'phones' or 'sub-phone' units were considered to be the ideal level of speech segmentation for unit selection, though contiguous sequences of phones were automatically

preferred as ‘non-uniform units’ by the original software. The phone-sized units are optimal for generation of novel words, particularly for use in ‘well-formed’ utterances, but in conversational speech, many of the utterances are NOT well-formed and many of the ‘words’ would not be found in a conventional dictionary. The ‘grunts’ of social interaction spoken merely for the sake of just ‘hanging-out’ are of a different order from the linguistic sounds of task-based speech (Trouvain & Truong, 2012; Gilmartin et al., 2018).

Furthermore, when the source data are virtually infinite (figuratively speaking) then the smallest speech segment might no longer be considered as optimal for concatenation, and a statistically-derived ‘non-linguistic’ chunk may be preferred instead (Deligne& Bimbot, 1997), more realistically reflecting the learnt patterns of speech behaviour (coarticulated speech gestures). There may always be a need for phonebased synthesis for novel words, but from a large corpus it is likely that many entire utterances or substantial portions of utterances can be reused intact. The task then is to index them in such a way that they can be rapidly selected for re-use in a novel utterance context.

5.1 Text Processing

Table 2 shows sample multigram units (and their pronunciation dictionary for ASRbased segmental re-alignment to the speech waveforms) that were automatically derived from the transcriptions by the processing illustrated in steps a1 a5 in Figure 2 above. They represent common idiomatic or colloquial phrasal chunks. Table 3 shows a sample of their bigram probabilities as calculated by the SRI Language Modelling toolkit (software). Kakasi (software) was used for the kanji/kanato-romaji conversion, and the romaji symbols have a direct mapping to the phonetic representations of Japanese speech sounds. Readers familiar with Japanese might be surprised by the highly colloquial nature of the resulting units and the preservation of the Kansai dialect speech forms in the utterance chunks.

Table 2. Multigrams derived from the transcribed corpus

N	N NNN	N N	
N NNNN	N N N N Nchau		
N ch a u NchauN	N ch a u		
N Nchauka	N ch a u k a		
Nchaukana	N ch a u k a n a		
Nchauno	N ch a u n o Nchauq		
N ch a u q			
Nde	N d e		
Nkai	N k a i		
NkamoshireNkedo N k a m o sh i r e N k e d o			
Nkana	N k a n a Nkanaa	N k a n	
a a Nkanaatoomoqte	N k a n a a t o o m o q t		
e Nkaq	N k a q		

Table 3. Example statistical language model probabilities for the multigram units

-0.001660784 uNN < /s >
-0.001693158 tomodachitonodeNwa < /s >
-0.001761846 NneNkedona < /s >
-0.001867936 teNyaN < /s >
-0.001884144 NneyaN < /s >
-0.001884144 maanaa < /s >
-0.001900635 hoNmaa < /s >
-0.00199676 teNkedona < /s >
-0.002024687 yaqteNyaN < /s >
-0.00203417 yawa < /s >
-0.002092989 < s > uNN < /s >
-0.002103125 tomodachitonokaiwa < /s >
-0.002134129 NneNyaN < /s >

5.2 Conversational Speech Unit Selection

As we proposed after preliminary work in 'User Interface for an Expressive Speech Synthesiser' (Campbell, 2004), the content and speaking style of an utterance may be realised as the expression of a discourse 'event' (E*) taking place within a framework of 'mood and interest' constraints (S for 'self') under 'friend or friendly?' restrictions (O for 'other'); i.e., U = E|(S,O) where S(sel f) represents the speaker's mood, interest, and +/engagement in the conversation, and O(other) represents a +/friendly partner and +/friendly intention towards the interlocutor.

"If motivation or interest in the content of the utterance is high, then the speech is typically more expressive. If the speaker is in a good mood then more so ... If the listener (other) is a friend, then the speech is typically more relaxed, and if in a friendly situation, then even more so . . ." (ibid).

The E'event0 was at that time considered to be primarily of either I-type (expressing 'information') or Atype (expressing 'affect'). This is clearly an over-simplification of the ideal case, but it remains worthy of testing and extending as an approximation, and as new understanding is gained from corpus analyses. Of particular interest of course, are the utterances which come under both categories, and a knowledge of how the combination is expressed through modulation of the voice or choice of expression is needed (Trouvain & Truong, 2012).

The framework described in Figure 4 and in the text cited above provides a means of using the higher-level features annotated in the GBO corpus directly for unit selection in the synthesis. An implementation was tested many years ago using an iMode (NTT) telephone interface but the response time was too slow. Now we have real-time interactive dialogue systems in which to test it, but the newer implementation using the entire corpus is currently still work in progress.

5.3 Input for Conversational Speech Synthesis

Whereas input for CHATR was in the form of written text (text-to-speech), the input for selection from a massive conversational speech corpus is necessarily

more complex. In addition to some way of specifying the text of each utterance, we also need to specify its purpose and information about its discourse contexts.

The 'text' may in fact be the least important aspect of a conversational utterance; consider for example the pragmatics of a simple morning greeting if to a close friend or family member, it may be just a simple 'grunt', but if to a stranger or work colleague then it may have a more formal aspect. The choice of 'words' is in fact less important than the 'expression' and 'tone-of-voice' in the speech.

For an interactive spoken dialogue system, there will be considerable contextual information available for such a choice to be made: Is the customer a first-time caller, or a regular interactant? Is the task a simple one or does it require more patient explanation? Is the call task-based, or 'merely' social?

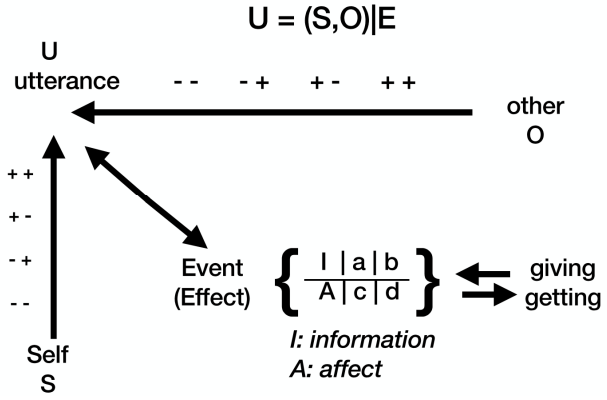


Figure 4.
Criteria for higher-level feature-based unit selection incorporating pragmatic constraints (from: Campbell (2004) User Interface for an Expressive Speech Synthesiser, IEICE Tech Rept.)

6 Conclusion

This chapter has described the series of steps that we are taking to process the GBO corpus for conversational speech synthesis using unit selection. We have succeeded in creating a unit database from a speech corpus and we now have a clearer understanding of the selection criteria that are needed to express a conversational utterance using natural speech in concatenative synthesis.

There still remains much work to be done in understanding the factors involved in non-task-based social interaction, and in how the voice is used in care-giving or informal friendly interactions, but we are confident that our corpus will provide the necessary answers through the processing described above.

Acknowledgements

This work has been made possible through the support of Japan Science & Technology Agency (for collection of the corpus) and Science Foundation Ireland (for funding of the research infrastructure in Ireland). The author is grateful to the School of Computer Science and Statistics and the ADAPT Centre in TCD for their kind encouragement of the continuing research.

References

Abadi, M., Agarwal, A., Barham, P. et al. (2016). *TensorFlow: Large-scale machine learning on heterogeneous distributed systems.* arXiv preprint arXiv:1603.04467, November, 2-4, 2016, Savannah, Georgia, USA.

Arik, S. O., Chrzanowski, M., Coates, A. et al. (2017). *Deep voice: Real-time neural text-to-speech.* arXiv preprint arXiv:1702.07825. August 6-11, 2017, Sydney, Australia.

Campbell, N. (1994). Prosody and the selection of units for concatenation synthesis. In *Proceedings of ESCA/IEEE 2nd w/s on Speech Synthesis* (pp. 61-64). Mohonk, N.Y. September 12-15, 1994, New York, USA.

Campbell, N. (1996). CHATR: A High-Definition Speech Re-sequencing System. In *Proceedings of ASA/ASJ Joint Meeting* (pp. 1223-1228). December 23-28, 1996, Hawaii, USA.

Campbell, N. (2001). Building a Corpus of Natural Speech and Tools for the Processing of Expressive Speech the JST CREST ESP Project. In *Proceedings of Interspeech 2001* (pp. 1525-1528). September 3-7, 2001, Aalborg, Denmark.

Campbell, N. (2004) User Interface for an Expressive Speech Synthesiser. *IEICE Tech Rept.,* 253-254.

Capes, T., Coles, P., Conkie, A. et al. (Apple Inc), (2017). Siri On-Device Deep Learning-Guided Unit Selection Text-to-Speech System. Interspeech August 20–24, 2017, Stockholm, Sweden

Deligne, S. (1996). Language Modeling By Variable Length Sequences

Deligne, S., & Bimbot, F., (1997). Inference of variable-length linguistic and acoustic units by multigrams. *Speech Communication, 23*(3), 223-241. https://doi.org/10.1016/S0167-6393(97)00048-4

Gilmartin, E., Spillane, B., Saam, Chr., Vogel, C., Campbell, N., & Wade, V. (2018). Stitching together the conversation considerations in the design of extended social talk. In Proceedings of IWSDS. May 14-16, 2018, Huone, Singapore.

Hunt, A., & Black, A. (1996). Unit selection in a concatenative speech synthesis system using a large speech database. In *Proceedings of ICASSP* (pp. 373-376). May 7-10, 1996, Atlanta, Georgia, USA.

KAKASI *Kanji Kana Simple Inverter,* (software) http://kakasi.namazu.org

Moulines, E., & Charpentier, F. (1990). Pitch synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Commnunication, 9,* 453-467.

Pollet, V., Zovato, E., Irhimeh, S., & Batzu, P. (Nuance Communications), (2017). *Unit selection with Hierarchical Cascaded Long Short Term Memory Bidirectional Recurrent Neural Nets.* Interspeech. August 20-24, 2017, Stockholm, Sweden.

Sagisaka, Y., Kaiki, N., Iwahashi, N., & Mimura, K. (1992). ATR v-talk speech synthesis system. In *Proceedings of ICSLP* (pp. 483-486). October 12-16, 1992, Banff, Alberta, Canada.

Sorin, A., Shechtman, S., Rendeli, A., (IBM), (2017). Semi Parametric Concatenative TTS with Instant Voice Modification Capabilities, Interspeech 2017 August 20-24, 2017, Stockholm, Sweden

SRILM https://www.sri.com/engage/products-solutions/sri-language-modeling-toolkit (software)

Trouvain, J., & Truong, K. (2012). Comparing non-verbal vocalisations in conversational speech corpora. In *Proceedings of the 4th International Workshop on Corpora for Research on Emotion Sentiment and Social Signals (ES3 2012)* (pp. 36-39). Paris, France: European Language Resources Association (ELRA).

Wan, V., Agiomyrgiannakis, Y., Silen, H., & Vit, J. (2017) Google's Next-Generation Real-Time Unit-Selection Synthesizer using Sequence-To-Sequence LSTM-based Autoencoders. Interspeech August 20-24, 2017, Stockholm, Sweden.

Wang, W, Xu, S, & Xu, B. (2016). First step towards end-to-end parametric TTS synthesis: Generating spectral parameters with neural attention. In *Proceedings Interspeech* (pp. 2243-2247). September 8-12, 2016, San Francisco, California.

DURATIONAL PATTERNS AND FUNCTIONS OF DISFLUENT WORD-REPETITIONS: THE EFFECT OF AGE AND SPEECH TASK

Judit BÓNA & Tímea VAKULA
ELTE Eötvös Loránd University, Budapest, Hungary
bona.judit@btk.elte.hu, vakula.timi@gmail.com

Abstract
The aim of this study is to analyse durational patterns and functions of disfluent whole-word repetitions in diverse age groups and speech tasks. Speech samples of school children (9-year-olds), adolescents (13-14-year-olds), young adults (20-25-year-olds) and old speakers (75+) were selected for the analysis. Recordings were made with each subject in two situations representing different speech tasks: 1) spontaneous narrative (participants spoke about their own lives and families), and 2) narrative recall (the task was to recall two texts they had listened to as accurately as possible). Results show that there are differences in the durational patterns and functions between the age groups in both speech tasks. Editing phases were significantly longer in 9-year-olds than in adults. In the ratio of the duration of R2 and R1, there were significant differences between 9-year-olds and the other three age groups, and between adolescents and the old speakers. As regards functions, in spontaneous narratives, the ratio of canonical repetitions was higher in 13- and 20-30-year-olds, and the ratio of stalling repetitions was higher in 9- and 75+-year-olds. In narrative recall, the ratio of stalling repetitions rose in 20-30- and 75+-year-olds. However, there were no significant differences between the speech tasks in any age group.

Keywords: whole-word repetition, durational patterns, function, speakers' age, speech task

1 Introduction

In spontaneous speech, one of the most frequent types of disfluency is word-repetition (Shriberg, 1995) that may stem from word-finding problems, difficulties in conceptual planning, or covert self-monitoring (Plauché & Shriberg, 1999). Word repetitions are considered disfluencies when the repeated word occurs due to speech planning and production problems. A repeated word is not considered disfluent when it occurs intentionally for emphasis or for pragmatic or stylistic reasons (Lickley, 2015). The differentiation of the two types of repetition (disfluency or pragmatic/stylistic role) is also supported by context and suprasegmental structure (intonation, speech rate and/or emphasis).

DOI: http://doi.org/10.18135/CAPSS.169

Repetitions consist of several parts: the original utterance, the first instance of the repeated word (R1), the second instance of the repeated word (R2) and the continuation of the utterance. Optional pauses may also occur next to the main parts: before the first instance (P1), between the two instances (P2, editing phase) and after the second instance (P3) (Plauché and Shriberg 1999). Example (1) shows the main parts of a disfluent whole-word repetition (SIL = silent pause):

(1) There is a book SIL on SIL on SIL the table.
Original utterance P1 R1 P2 R2 P3 Continuation

The phonetic characteristics of R1 and R2 were analysed in several studies (Shriberg, 1999, 2001; Gyarmathy, 2009; Bóna, 2010). It was found that as regards durations, R1 and R2 can be realised in three different ways: (i) R1 is longer than R2; (ii) R2 is longer than R1; or (iii) the duration of R1 and R2 is similar. The last case is quite rare while the first one is the most frequent. For example, in English, repetitions of the article *the* were analysed. In this case, R1 was significantly longer than R2. The duration of R2 was similar to the duration of the article occurring in fluent speech (Shriberg, 1999). Based on these data, it was concluded that speakers try to avoid silent or filled pauses. They make an effort to keep their speech fluent by lengthening R1 (Shriberg, 1999). In Gyarmathy's study (2009), R1 was longer in 71.95% of all repetitions. Considering all repetitions, there was significant difference in duration between R1 and R2. In addition to duration, f0 and formants of vowels were also analysed (Shriberg, 1999; Gyarmathy, 2009). Results showed that there was no significant difference in these parameters. This proves that R1 and R2 are parts of a single phonetic plan (Gyarmathy, 2009).

The duration of the first and second instances of the repeated words and the occurrence of pauses are related to the function of the repetitions. Heike (1981) defines two functions of disfluent whole-word repetitions: (i) R2 is the hesitation in itself, in other words, it fills the gap caused by speech planning problems (prospective repeats); (ii) R2 is the bridge between original utterance and continuation (retrospective repeats). In this case, planning difficulties are solved during the pronunciation of R1. The two functions are characterized by pauses occurring before, between and after R1 and R2. In the first case, R2 is followed by a pause. In the second case, R2 is preceded by a pause but it is not followed by one. According to Shriberg (1995), the second type of repetitions is significantly more frequent than the first type. The duration of R1 and R2 and their ratio depend on the function of whole-word repetitions. If the repetition is prospective, R2 is longer than R1. If the repetition is retrospective, R1 is longer than R2. Plauché and Shriberg (1999) found three main types of functions: canonical repetition, covert self-repair, and stalling repetition (Table 1). Their

categorization was based on the durational patterns of word-repetitions, but f0 variation and glottalization were also considered. The categorization of Plauché and Shriberg (1999) could be valid for any language, although they examined only *I* and *the*. Irrespective of which words are considered disfluent repetitions, distinctions can be made between the different functions.

In cases of canonical repetition (Plauché & Shriberg, 1999), the duration of R1 is much longer than in the utterance of the same word in fluent speech. The duration of R2 is similar to the fluent word. There might be a pause before R1, there is a long pause between R1 and R2, and there is no pause after R2. Both R1 and R2 are characterized by falling intonation, and R1 is often characterized by diplophonia and creak-like voicing modality (similar to a filled pause). In this case the speaker has difficulties during speech production, stops during the pronunciation of the word (R1), lengthens it, and after having solved the problem they continue speaking with repeating the last lengthened word. This type corresponds to Heike's retrospective repeat (1981).

In cases of covert self-repair (Plauché & Shriberg, 1999), P1 often occurs, but there is no P2 or P3. R1 and R2 are slightly longer than they are in fluent speech, and their durations are similar to each other. R1 and R2 are both characterized by rising pitch. R1 is sometimes pronounced with glottalization. In this case the speaker detects a problem during the pronunciation of R1; this is shown by a possible preceding pause and glottalization. The speaker makes an effort to correct it, and "R2 usually marks the beginning of a new utterance or a corrected version of the previous one" (Plauché & Shriberg, 1999, p. 1516).

In cases of stalling repetition (Plauché & Shriberg, 1999), there is no pause before R1, but P2 and P3 may occur. The duration of R1 is slightly longer than in fluent speech, and the duration of R2 is much longer. R1 is characterized by a drop in pitch. The speech is fluent during the pronunciation of R1, the speaker has a problem during and/or after the pronunciation of R2. This is usually marked by P3 or other possible disfluencies after R2. This type looks as if it was the inverse of canonical repetitions, and corresponds to Heike's prospective repeat (1981).

The categories of Plauché and Shriberg (1999) are determined by hierarchical clustering based on acoustic data. Out of the 819 whole-word repetitions analysed, 724 were distributed in these main categories. The remaining 95 occurrences were distributed across 32 other clusters.

The characteristics of repetitions (like other disfluencies) are influenced by several factors: for example, by the age of the speaker. DeJoy and Gregory (1985) found that in 3.5- and 5-year-old children's speech one of the most frequent disfluencies is whole-word repetition. 3.5-year-old children produce word-repetitions significantly more frequently than 5-year-olds. Similar results were found by Kowal et al. (1975). They found that the occurrence of word-

repetitions fell to one-sixth between preschool- and secondary-school-age. In the speech of 6-7-year-old Hungarian speaking children, the ratio of word-repetitions was the highest (43%) among all disfluencies (it was even higher than the ratio of filled pauses – the latter was 16%) (Horváth, 2006). According to Neuberger (2014), word-repetition was the second most frequent disfluency-type in 6-year-olds' speech. However, above age 7, its ratio was only 3-14%. Bóna (2013) analysed word-repetitions in old speakers' speech. Her results show that the occurrence of word-repetitions is significantly less frequent in old speakers' speech than in young speakers' speech. Editing phases (P2) of old speakers were significantly shorter than those of young speakers. The ratio of zero editing phases was higher in old speakers' speech. Bóna and Vakula (2017) found that whole-word repetitions of content words are more frequent in children's and old speakers' speech and the occurrence of stalling repetition is more frequent in their case compared to young and middle-aged adults.

Table 1. The structures of the three types of word-repetitions (examples with 'the') '+' = a longer than fluent duration. ' - ' = no pause (based on Plauché and Shriberg 1999)

Type	Structure
Canonical repetition	(Original Utterance) (Possible Pause) the+++ (Long Pause) the (-) (Continuation)
Covert self-repair	(Original Utterance) (Often Pause) the+ (-) the+ (-) (Continuation)
Stalling repetition	(Original Utterance) (-) the+ (Possible Pause) the+++ (Possible Pause) (Continuation)

Speakers' age influences not only the frequency of disfluencies, but also temporal characteristics of speech. As children are getting older, speech rate accelerates, although this acceleration is non-linear (Walker & Archibald, 2006). The change of speech rate and articulation rate happens due to biological factors and learned skills. Biological factors are the neurologic and neuromotor maturation (Smith et al., 1983; Smith, 1992); learned skills are motor learning, semantic, lexical, phonological access, and motor programming and planning (Nip & Green, 2013; Redford, 2014). Working-memory performance and speech rate are also related to each other: the speech rate of older children is positively influenced by the increase in storage capacity of working-memory and the better functioning of long-term memory (Roodenrys et al., 1993; Henry, 1994).

Speech rate becomes slower in the elderly (e.g., Hartman & Danhauer, 1976; Ramig, 1983; Duchin & Mysak, 1987; Bóna, 2014). There are several reasons in the background of the differences in speech rate of speakers of different ages: hormonal, psychological, and cognitive changes (Rodríguez-Aranda & Jakobsen, 2011); the aging of the speech organs (Xue & Hao, 2003), and the deterioration of hearing (Chisolm et al., 2003). Durational patterns of

disfluencies (such as repetitions) might be affected by all of these age-related changes in speech and articulation rates.

In addition to age, the speech task presumably also influences the occurrence of whole-word repetitions. This is due to the fact that the different speech tasks require different speech planning mechanisms. The differences of speech planning mechanisms in these speech tasks show up in temporal characteristics (the frequency and duration of pauses can show speech planning processes), too (Ramig, 1983; Duchin & Mysak, 1987; Jacewicz et al., 2010; Bóna, 2014; Redford, 2015). Comparisons of narratives and conversations show that there is significant difference in the frequency of disfluencies between the two speech tasks (Shriberg, 2005; Beke et al., 2014). Furthermore, there are differences in the occurrences of disfluencies between narratives with different topics (Roberts et al., 2009). In an analysis of Hungarian speech, the ratio of whole-word repetitions within all disfluencies was different in various speech tasks. The most frequent occurrence was found in narrative recall, and the less frequent in spontaneous narratives (Bóna, 2014).

The question is how speakers' age and the speech task influence the durational patterns and functions of disfluent whole-word repetitions. The aim of this study is to analyse durational patterns and functions of the repeated words in diverse age groups and speech tasks. The hypotheses of the research are: (i) there will be a difference in the durational patterns and functions of repetitions between the age groups in both speech tasks; (ii) there will be a significant difference between the speech tasks in the characteristics of repetitions in each age group.

2 Methods

For the analysis, speech recordings of 80 speakers were selected from two Hungarian speech databases. Speech samples of schoolchildren (9-year-olds), and adolescents (13-14-year-olds) were selected from GABI Hungarian Children Speech Database and Information Repository (Bóna et al., 2014). Speech samples of young adults (20-25-year-olds) and old speakers (75+) were selected from BEA Hungarian Speech Database (Gósy, 2012). In every age group there were 20 speakers (10 females and 10 males). They were native Hungarian speakers with normal hearing and without any known mental or speech disorders. They spoke standard Hungarian.

Recordings were made with each subject in two situations which represented different speech tasks: (i) spontaneous narrative, and (ii) narrative recall. In spontaneous narratives participants spoke about their own lives and families. They could speak freely and use words and grammatical forms of their own choice. In narrative recall, the task was to recall two texts they had listened to as accurately as possible. One was a science dissemination text, the other was a

historical anecdote. The texts were the same in each age group. In this speech task, the success of recalls was determined by speech processing, attentional and working memory mechanisms, and narrative competence (Juncos-Rabadán and Pereiro 1999). Altogether about 8 hours of speech were analysed.

Disfluent whole-word repetitions were collected from the recordings. The analysis was not aimed at determining frequency of occurrence so instances per person were not calculated. Altogether 446 whole-word repetitions were analysed. The number of occurrences of whole-word repetitions depending on age and speech task is shown in Table 2.

Table 2. Number of occurrence of whole-word repetitions depending on age and speech task

	Spontaneous narratives	**Narrative recalls**
9-year-olds	32	17
13-year-olds	21	15
20-30-year-olds	160	57
75+-year-olds	94	50
All	307	139

The annotations and measurements (duration of the components of repetitions) were carried out by Praat (Figure 1). The first and second instances of the repeated word and the pauses between them were measured. We analysed the ratio of the second (R2) and the first (R1) instance of the repeated word and the pauses between and after them. This was needed because differently from Plauché and Shriberg (1999), in this analysis every disfluently repeated word was examined. This means that not only *én* 'I' and the definite article *a, az* 'the' were analysed, but also other disfluently repeated function words (e.g. conjunctions) or content words. The difficulty was caused by the fact that participants did not repeat the same words in each age group and in both speech tasks. So the comparison of the raw duration of R1 and R2 was not possible. In addition, this method (comparing the ratio of R2 and R1) also allowed us to eliminate the influence of the differences in articulation rate.

Pauses before the first instance and after the second instance of the repeated words were not measured, but their occurrences were considered for the examination of functions. Functions were determined on the basis of the durational patterns of R1 and R2, and the occurrence of pauses. The categorization was supported also by the perceived intonation (falling or rising) and voicing features (glottalized or not) (based on Plauché & Shriberg, 1999; see Table 1). According to the above types, the analysed whole-word repetitions were categorized in four groups: (i) canonical repetitions, (ii) covert self-repairs, (iii) stalling repetitions, (iv) other (the cases which could not be categorized in the main types). The group of multiple repetitions which contains cases where

the first instance of the word is repeated more than once was not analysed. Types of the repeated words (content word or function word) were also determined. The measurements and classifications were carried out by the two authors independently. After that, 10% of the data was reanalysed by the other author. The results of the two analyses were similar in 100% of the cases.

Statistical analysis (Wilcoxon Signed Ranks Test, Mann–Whitney-test, repeated measures ANOVA) was carried out by SPSS on 95% confidence level.

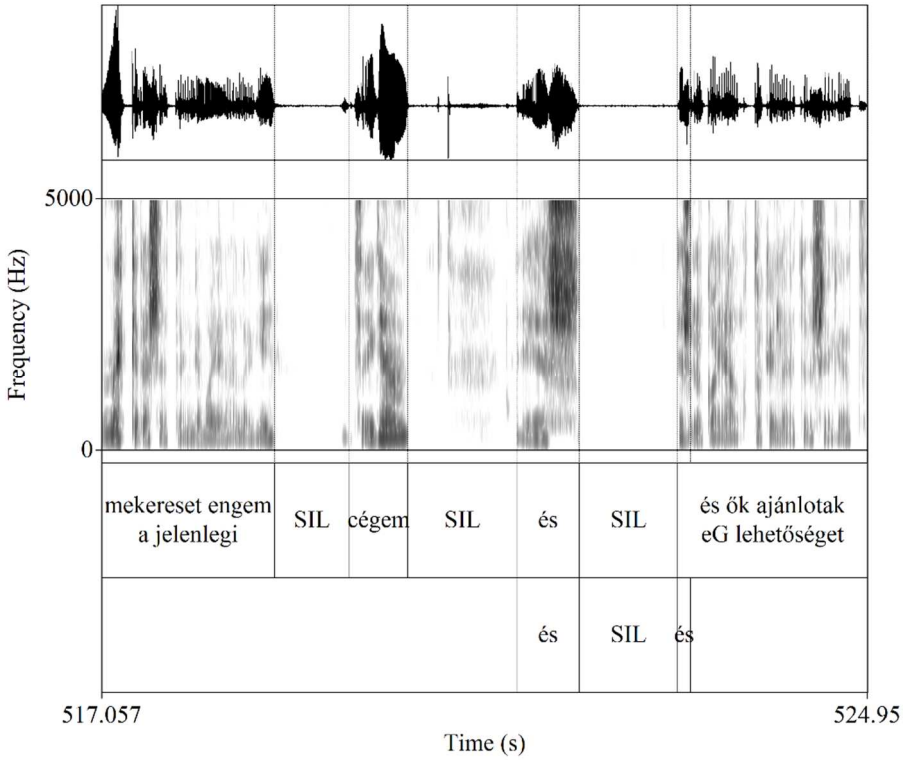


Figure 1.
Example for the annotation by Praat (SIL = silent pause)

3 Results

First, types of the repeated words were determined (Figure 2). In each age group, function words were repeated in a higher ratio than content words were. However, the ratio of repeated content words was much higher in 9-year-olds and 13-year-olds than in 20-30- and 75+-year-olds. In addition, 75+-year-olds produced twice as high a ratio of repetitions of content words than 20-30-year olds. In addition to age, speech tasks also influenced the ratio of the repetitions of content words. Their ratio was higher in spontaneous speech in all four age

groups. The biggest difference between the two speech tasks appeared in 9-year-olds: the appearance of repeated content words was 28.1% in spontaneous speech and 17.6% in narrative recall. The smallest difference between the two speech tasks appeared in young adults: it was only 0.25 percentage points. The ratio of repeated content words in the case of young adults was 3.75% in spontaneous speech and 3.5% in narrative recall.

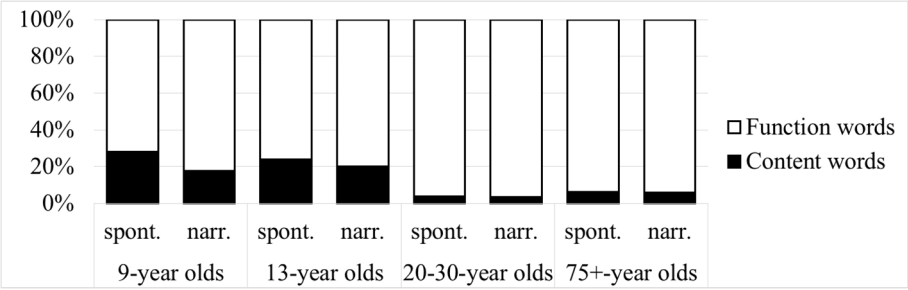


Figure 2.
The ratio of content words and function words
(spont. = spontaneous speech, narr. = narrative recall)

The duration of R1 and R2 was analysed in all repetitions (Table 3). In the case of 9-year-olds, there was no significant difference between R1 and R2 in spontaneous speech. However, there was significant difference in narrative recall [repeated measures ANOVA: $F(1, 16) = 15.673$, $p = 0.001$, $\eta^2 = 0.495$]. In the case of 13-year-olds, there were significant differences between the durations of R1 and R2 in spontaneous speech [repeated measures ANOVA: $F(1, 20) = 28.755$, $p < 0.001$, $\eta^2 = 0.590$] and in narrative recall (Wilcoxon Signed Ranks Test: $Z = -3.294$, $p = 0.001$). In the case of 20-30-year-olds, there were signify-cant differences between R1 and R2 in spontaneous speech [repeated measures ANOVA: $F(1, 159) = 91.871$, $p < 0.001$, $\eta^2 = 0.366$] and in narrative recall (Wilcoxon Signed Ranks Test $Z = -3.869$, $p < 0.001$). In the case of 75+-year-olds, there was no significant difference between R1 and R2 in any speech task.

Table 3. Duration of R1 and R2 depending on age and speech task (ms)
(Mean ± Standard Deviation)

	Spontaneous narratives		**Narrative recalls**	
	R1	**R2**	**R1**	**R2**
9-year-olds	430±199	378±212	582±256	418±189
13-year-olds	448±179	255±113	600±328	377±202
20-30-year-olds	344±144	232±109	349±149	268±185
75+-year-olds	362±237	334±181	265±159	272±200

To be able to compare how the duration of R1 and R2 relate to each other across age groups and speech tasks, the ratio of R2 and R1 was calculated (Figure 3). If the ratio was less than 100%, R1 was longer than R2. If the ratio was more than 100%, then R2 was longer than R1. In the case of adolescents and young adults, the majority of the values were below 100%. In the case of schoolchildren and old speakers, the majority of the values were over 100%. The ratio of R2 and R1 was 92±9.1% in schoolchildren's spontaneous narratives, and 74±5.2% in their narrative recalls. It was 63±6.5% in adolescents' spontaneous narratives, 67±5.3% in their narrative recalls. 75±3.2% in young adults' spontaneous narratives, 80±6.8% in their narrative recalls. 107±6.3% in the old speakers' spontaneous narratives, 110±7.1% in their narrative recalls. Table 4 shows the significant differences as results of the statistical analysis. There were no significant differences between 20-30-year-olds and 13-year-olds, and between 20-30-year-olds and 75+-year-olds, in any speech tasks.

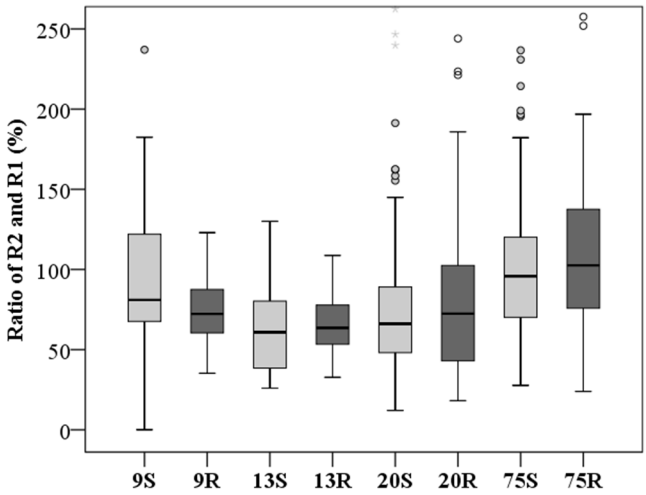


Figure 3.
Ratio of the durations of R2 and R1
(R2 = duration of the second instance of the repeated word, R1 = duration of the first instance of the repeated word, S = spontaneous speech, R = narrative recall)

Table 4. Significant differences between the age groups in the ratio of R2 and R1 (Results of the Mann–Whitney-test)

	Spontaneous narratives		**Narrative recalls**	
	Z	***p***	***Z***	***p***
9- and 13-year-olds	−2.237	0.025	–	–
9- and 20-30-year-olds	−2.805	0.005	−2.531	0.011
9- and 75+-year-olds	−2.365	0.018	−2.662	0.008
13- and 75+-year-olds	−3.866	< 0.001	−3.519	< 0.001

Editing phases (P2) of all repetitions were also analysed (Figure 4). The longest editing phases were produced by 9-year-olds. According to the statistical analysis, there were significant differences between 9-year-olds and 20-30-year-olds (Mann–Whitney-test, spontaneous narratives: $Z = -2.805$, $p = 0.005$; narrative recalls: $Z = -2.531$, $p = 0.011$) and between 9-year-olds and 75+-year-olds (Mann–Whitney-test, spontaneous narratives: $Z = -2.365$, $p = 0.018$; narrative recalls: $Z = -2.662$, $p = 0.008$) in the duration of editing phases in the two speech tasks. There was no significant difference between the two speech tasks in any of the age groups.

The majority of editing phases was realized as silent pause in each age group and in both speech tasks. In 20-30-year-olds and 75+-year-olds, the ratio of silent editing phases was higher than in the other two age groups (Figure 5).

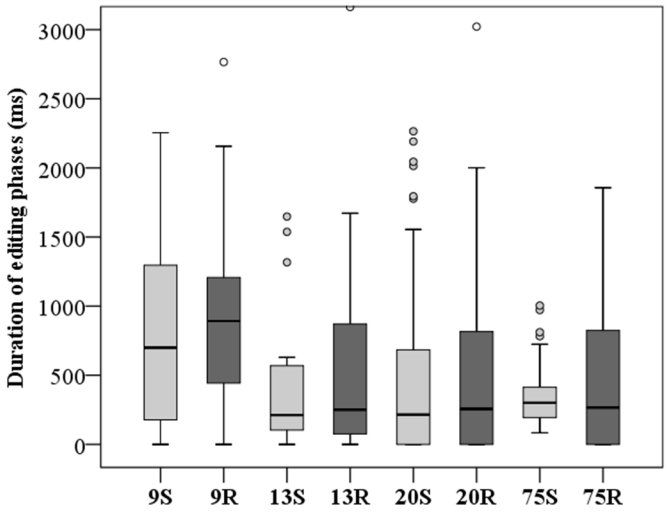


Figure 4.
Duration of editing phases of every repetition
(S = spontaneous speech, R = narrative recall)

The functions of repetitions were also analysed (Figure 6). Schoolchildren and adolescents produced canonical repetitions in higher ratio than the other two age groups. In the old speakers' speech, stalling repetitions and "other types" occurred in a much higher ratio than in the other groups. Covert self-repairs occurred in the highest ratio in 20-30-year-olds' spontaneous narratives. In the comparison of speech tasks, the ratio of canonical repetitions was higher, and the ratio of covert self-repairs was lower in narrative recall than in spontaneous narrative, in each age group.

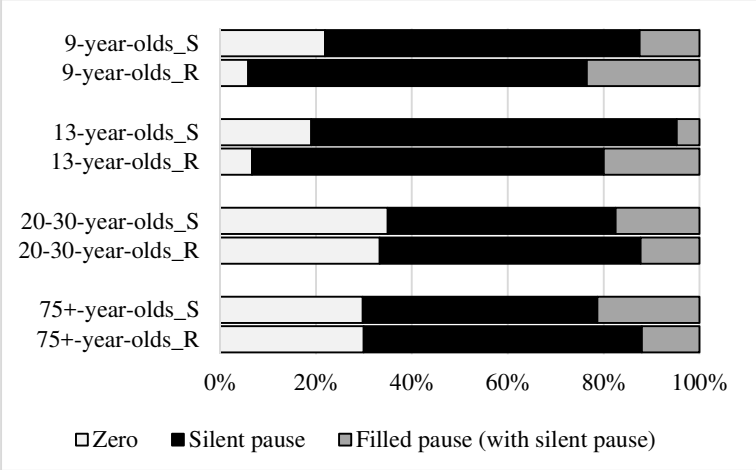


Figure 5.
Types of editing phases (P2)
(S = spontaneous speech, R = narrative recall)

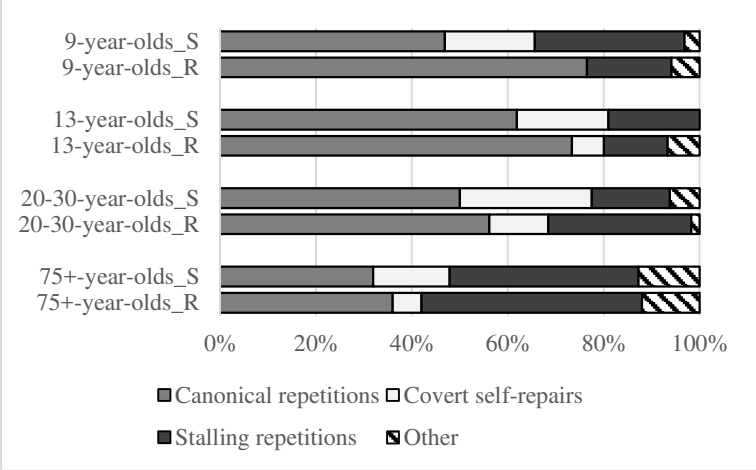


Figure 6.
Functions of repetitions
(S = spontaneous speech, R = narrative recall)

The ratio of R2 and R1 and editing phases (P2) depending on types of functions of repetitions were also analysed. According to the statistical analysis, in cases of covert self-monitoring, there were no significant differences in the ratio of R2 and R1 between any age groups and any speech tasks (in this case, the editing phase was always 0 ms).

Results of canonical repetitions are shown in Table 5. According to the statistical analysis, in spontaneous narratives, there was significant difference in the ratio of R2 and R1 between 20-30-year-olds and 75+-year-olds (UniANOVA: $F(3, 134) = 4.268$, $p = 0.007$; $\eta^2 = 0.087$; Tukey's post hoc test: $p = 0.003$). In narrative recalls, there was a significant difference also between 20-30-year-olds and 75+-year-olds in the ratio of R2 and R1 (UniANOVA: $F(3, 71) = 5.329$, $p = 0.002$, $\eta^2 = 0.186$; Tukey's post hoc test: $p = 0.006$). There were no significant differences between the two speech tasks.

Table 5. Ratio of R2 and R1 and duration of editing phases of canonical repetitions (mean±SD)

	Ratio of R2 and R1 (%)		**Duration of editing phases (ms)**	
	Spontaneous narratives	**Narrative recalls**	**Spontaneous narratives**	**Narrative recalls**
9-year-olds	58±8.6	70±4.7	1421±313	1191±489
13-year-olds	53±6.4	68±7.2	630±148	318±114
20-30-year-olds	53±2.1	50±3.6	636±60	730±177
75+-year-olds	69±4.2	72±6.3	607±158	758±141

As regards editing phases, in spontaneous narratives, there were significant differences between 9-year-olds and 13-year-olds (Mann–Whitney-test: $Z = -2.326$, $p = 0.020$), 9-year-olds and 20-year-olds ($Z = -3.041$, $p = 0.002$), and 9-year-olds and 75+-year-olds ($Z = -3.226$, $p = 0.001$). In editing phases, in narrative recall, there were significant differences between 13-year-olds and 9-year-olds (Mann–Whitney-test: $Z = -2.231$, $p = 0.026$), and 13-year-olds and 75+-year-olds ($Z = -2.247$, $p = 0.025$). There was no significant difference between the two speech tasks.

In case of stalling repetitions, in 9-year-olds and 13-year-olds there were so few data available that these two age groups could not be included in the statistical analysis. There was a significant difference between 20-30-year-olds and 75+-year-olds only in the ratio of R2 and R1, and only in spontaneous narratives ($Z = -2.164$, $p = 0.030$). In editing phases, and between the two speech tasks, there were no significant differences in any age groups.

4 Discussion and conclusion

In this paper durational patterns and functions of disfluent whole-word repetitions were analysed in four age groups and in two speech tasks. Results show that children, adolescents and old speakers repeat content words in much higher ratios than young adults do. This might mean that the former have more serious word retrieval or speech-planning and monitoring problems than the latter. This assumption is supported by the fact that in narrative recall, the ratio of repetitions of content words was reduced. Namely, in the case of narrative recall, it is not the speaker who has to select the appropriate word from their

vocabulary, since they already heard the words and grammatical forms of the story before they were asked to retell it.

As regards durational patterns of all repetitions, there were significant differences between the age groups, but not between the speech tasks. In editing phases, there were significant differences between 9-year-olds and the two adult groups. It seems as if adolescents formed a transition between schoolchildren and adults in this respect. Editing phases were significantly longer in 9-year-olds than in adults. On the one hand, they might have needed more time for solving the planning difficulties. On the other hand, they might not have felt the need to fill the gap as soon as possible with pronouncing the second instance of the repeated word (R2) during solving the speech-planning difficulties. In the ratio of the duration of R2 and R1, there were significant differences between 9-year-olds and the other three age groups, and between adolescents and old speakers. The smallest ratio of the duration of R2 and R1 was in adolescents in both speech tasks. This means that they pronounced R2 much shorter compared to R1 than the other groups. The ratio of the duration of R2 and R1 was over 100% in the case of old speakers. This means that they pronounced R2 longer than R1.

These durational patterns show the differences between the groups in the distribution of functions of whole-word repetitions. In spontaneous narratives, the ratio of canonical repetitions was higher in 13- and 20-30-year-olds, and the ratio of stalling repetitions was higher in 9- and 75+-year-olds. This shows bigger speech-planning difficulties in the latter groups. In the comparison of speech tasks, it seems that in narrative recall, repetitions in the function of covert self-repair occurred rarely in each age group. In narrative recall, the ratio of stalling repetitions rose in 20-30- and 75+-year-olds. This means that they have more speech-planning problems in this speech task or they realize them later because it was necessary for them to remember the story. The rise of the ratio of canonical repetitions in 9- and 13-year-olds might be caused by the decrease of the ratio of covert self-repair in narrative recall.

In the comparison of the durational patterns depending on functions, the results were similar to the comparison of the durational patterns of all repetitions. In case of canonical repetitions, 9-year-olds' data were the most divergent from the other age groups. Covert self-repair was similar in each age group. In case of stalling repetitions, only the two adult groups were comparable and they were mostly similar. There were no significant differences between the speech tasks in the durational patterns in any of the three functions.

Results show that the first hypothesis was confirmed: there are differences in the durational patterns (duration of the repeated words and pauses) and functions between the age groups in both speech tasks. The second hypothesis was not confirmed: there was no significant difference between the speech tasks in any age group. The speech task can influence the ratio of the different functions of

disfluent word-repetitions, but the ratio of the repeated words and the duration of editing phases were similar in both speech tasks.

The differences between age groups in the distribution of functions of whole-word repetitions indicate that in different age groups not only the frequency of disfluencies, but also their functions may indicate different speech planning strategies and difficulties. It would be worth examining the functions (not only the frequency) of certain other types of disfluency, too.

Acknowledgements

The authors wish to thank Zsófia Koren-Dienes for their help in preparing this paper. This research was supported by the Hungarian National Research, Development and Innovation Office of Hungary, project No. K-120234.

References

Beke, A., Gósy, M., Horváth, V., Gyarmathy, D., & Neuberger, T. (2014). Disfluencies in Spontaneous Narratives and Conversations in Hungarian. In S. Fuchs, M. Grice, A. Hermes, L. Lancia, & D. Mücke (Eds.), *Proceedings of the 10th International Seminar on Speech Production (ISSP)* (pp. 29-32). http://www.issp2014.uni-koeln.de/wp-content/uploads/2014/Proceedings_ISSP_revised.pdf

Bóna, J. (2010). Bizonytalansági megakadások idősek és fiatalok spontán beszédében. [Disfluencies in young and elderly adults' spontaneous speech.] *Beszédkutatás 2010,* 125-138.

Bóna, J. (2013). *A spontán beszéd sajátosságai az időskorban* [Characteristics of spontaneous speech in the elderly.] Beszéd – Kutatás – Alkalmazás 2. Budapest: ELTE Eötvös Kiadó.

Bóna, J. (2014). Temporal characteristics of speech: The effect of age and speech style. *Journal of the Acoustical Society of America, 136*(2), EL116-EL121.

Bóna, J., Imre, A., Markó, A., Váradi, V., & Gósy, M. (2014). GABI – Gyermeknyelvi Beszédadatbázis és Információtár. [GABI – Child Speech and Information Database.] *Beszédkutatás 2014*, 246-252.

Bóna, J., & Vakula, T. (2017). *Disfluent word-repetitions across the lifespan.* Paper presented at the Workshop on Speech Perception and Production across the Lifespan. 26-27 April 2017. London, UK.

Chisolm, T. H., Willott, J. F., & Lister, J. J. (2003). The aging auditory system: anatomic and physiologic changes and implications for rehabilitation. *International Journal of Audiology, 42*, 2S3.

DeJoy, D. A., & Gregory, H. H. (1985). The relationship between age and frequency of disfluency in preschool children. *Journal of Fluency Disorders, 10*(2), 107-122.

Duchin, S. W., & Mysak, E. D. (1987). Disfluency and rate characteristics of young adult, middle-aged, and older males. *Journal of Communication Disorders, 20*, 245-257.

Gósy, M. (2012). BEA – A multifunctional Hungarian spoken language database. *The Phonetician, 105-106,* 50-61.

Gyarmathy, D. (2009). A beszélő bizonytalanságának jelzései: ismétlések és újraindítások. [Marks of uncertainty of speakers: whole-word repetitions and part-word repetitions.] *Beszédkutatás 2009,* 196-216.

Hartman, D. E., & Danhauer, J. L. (1976). Perceptual features of speech for males in four perceived age decades. *Journal of the Acoustical Society of America, 59*(3), 713-715.

Heike, A. E. (1981). A content-processing view of hesitation phenomena. *Language and Speech, 24*(2), 147-160.

Horváth, V. (2006). A spontán beszéd és beszédfeldolgozás összefüggései gyerekeknél. [Correlation between spontaneous speech and speech perception in children.] *Beszédkutatás 2006,* 134-146.

Jacewicz, E., Fox, R. A., & Wei, L. (2010). Between-speaker and within-speaker variation in speech tempo of American English. *Journal of the Acoustical Society of America, 128*, 839-850.

Juncos-Rabadán, O., & Pereiro, A. X. (1999). Telling stories in the elderly. Influence of attentional and working memory proceses (preliminary study). In M. da Graça Pinto, J. Veloso, & B. Maia (Eds.), *Psycholinguistics on the threshold of the year 2000. Proceedings of the 5th International Congress of the International Society of Applied Psycholinguistics* (pp. 155-159). Porto: Faculdade de Letras da Universidade do Porto.

Kowal, S., O'Connell, D. C., & Sabin, E. J. (1975). Development of temporal patterning and vocal hesitations in spontaneous narratives. *Journal of Psycholinguistic Research, 4*(3), 195-207.

Lickley, R. J. (2015). Fluency and disfluency. In M. A. Redford (Ed.), *The handbook of speech production* (pp. 445-474). Malden, MA: John Wiley & Sons, Inc.

Neuberger, T. (2014). *A spontán beszéd sajátosságai gyermekkorban* [Characteristics of children's spontaneous speech]. Budapest: ELTE Eötvös Kiadó.

Nip, I. S., & Green, J. R. (2013). Increases in cognitive and linguistic processing primarily account for increases in speaking rate with age. *Child development, 84*(4), 1324-1337.

Plauché, M., & Shriberg, E. (1999). Data-driven subclassification of disfluent repetitions based on prosodic features. In *Proceedings of the International Congress of Phonetic Sciences* (Vol. 2., pp. 1513-1516).

Ramig, L. A. (1983). Effects of physiological aging on speaking and reading rates. *Journal of communication disorders, 16*(3), 217-226.

Redford, M. A. (2014). The perceived clarity of children's speech varies as a function of their default articulation rate. *Journal of the Acoustical Society of America, 135*(5), 2952-2963.

Redford, M. A. (2015). The Acquisition of Temporal Patterns. In M. A. Redford (Ed.), *The Handbook of Speech Production* (pp. 379-403). Malden, MA: John Wiley & Sons, Inc.

Roberts, P. M., Meltzer, A., & Wilding, J. (2009). Disfluencies in non-stuttering adults across sample lengths and topics. *Journal of Communication Disorders, 42*(6), 414-427.

Rodríguez-Aranda, C., & Jakobsen, M. (2011). Differential contribution of cognitive and psychomotor functions to the age-related slowing of speech production. *Journal of the International Neuropsychological Society, 17*(5), 807-821.

Roodenrys, S., Hulme, C., & Brown, G. (1993). The development of short-term memory span: Separable effects of speech rate and long-term memory. *Journal of Experimental Child Psychology, 56,* 431-431.

Shriberg, E. (1995). Acoustic properties of disfluent repetitions. In *Proceedings of the International Congress of Phonetic Sciences* (Vol. 4., pp. 384-387).

Shriberg, E. (1999). Phonetic consequences of speech disfuency. In *Proceedings of the International Congress of Phonetic Sciences.* (Vol. 1., pp. 619-622).

Shriberg, E. (2001). To 'errrr' is human: ecology and acoustics of speech disfluencies. *Journal of the International Phonetic Association, 31*(1), 153-169.

Shriberg, E. (2005). Spontaneous speech: How people really talk and why engineers should care. In *Proceedings of Interspeech* (pp. 1781-1784).

Smith, B. L. (1992). Relationships between duration and temporal variability in children's speech. *Journal of the Acoustical Society of America, 91*(4), 2165-2174.

Smith, B. L., Sugarman, M. D., & Long, S. H. (1983). Experimental manipulation of speaking rate for studying temporal variability in children's speech. *Journal of the Acoustical Society of America, 74*(3), 744-749.

Walker, J. F., & Archibald, L. M. (2006). Articulation rate in preschool children: a 3-year longitudinal study. *International Journal of Language & Communication Disorders, 41*(5), 541-565.

Xue, S. A., & Hao, G. J. (2003). Changes in the human vocal tract due to aging and the acoustic correlates of speech production: a pilot study. *Journal of Speech, Language, and Hearing Research, 46*(3), 689-701.

SPEAKER AGE ESTIMATION BY MUSICIANS AND NON-MUSICIANS

Ákos GOCSÁL
University of Pécs. Faculty of Music and Visual Arts, Institute of Music &
Hungarian Academy of Sciences, Research Institute for Linguistics
gocsal.akos@pte.hu

Abstract
Speaker age estimation is one of the most commonly researched fields in the domain of social perception based on voice. Previous findings confirm a strong correlation between the estimated and calendar age of speakers, however, younger adult speakers are usually perceived to be older, while older speakers are thought to be younger than their actual age. Effects of listener factors, such as age and gender have also been researched. The purpose of the present study is to examine if a more sophisticated auditory mechanism, which can be attributed to music training, results in more accuracy in speaker age estimation. The present research found correlation coefficients between calendar ages and mean estimated ages comparable to those reported in the literature, and musicianship and listener gender were not proven to have a significant effect on age estimations. Linear mixed models, implemented on three age groups, revealed some marginal differences between musicians and non-musicians, implying musicians' more accurate age estimations in some cases.

Keywords: speaker age estimation, social perception, musicianship

1 Introduction

1.1 Social perception

When we hear another person speaking, it is not only the linguistic message that we decode from the acoustic structures of speech. As Krauss (2002) says, in addition to what is said, human voice conveys considerable information about the speaker, and listeners use this information in human interaction. Based on speech, we infer a variety of speaker traits and make social judgments. These judgments are, however, based on vocal stereotypes and may even lead to prejudices (Drager, 2010). While a wide range of experiments have been conducted to explore humans' ability to infer different objective traits, such as speaker gender (Gelfer & Bennett, 2014), body size parameters (Rendall et al., 2007; van Dommelen & Moxness, 1995), or age (Moyse et al., 2014), a growing body of literature deals with the formation of voice-based impressions and

DOI: http://doi.org/10.18135/CAPSS.185

attitudes, including pleasantness (Hughes & Harrisson, 2013), attractiveness (Feinberg et al., 2005; Abend et al., 2015), perceived dominance (Puts et al., 2006; Fraccaro et al., 2013) or competence (Klofstad et al., 2015). Also, the perception of accented language or a dialect may evoke certain attributions and attitudes towards the speaker (Rubin et al., 1991, Cargile & Giles, 1997, Cross et al., 2001). Such impressions and attitudes may serve as bases for further decisions and actions; for example, voting behavior (Klofstad et al., 2015; Tigue et al., 2012) or mate choice (Collins, 2000; Shoup-Knox & Pipitone, 2015) may be influenced.

Little is known, however, about whether the perceptual behavior of individuals with different characteristics, such as age, gender, or any other factor, differs in their judgments. The results obtained so far seem to be equivocal. For example, Rendall et al. (2007) found no significant differences between male and female listeners' speaker size judgments; however, in an experiment by Charlton et al. (2013), men were better at classifying the apparent size of stimuli than female participants. In a similar experiment, Pisanski et al. (2016) did not find significant differences between sighted and congenitally blind subjects, and those who lost their sight later in life; and they did not find gender differences, either.

One factor that may differentiate listeners in their judgments is the difference in their auditory skills. In the context outlined above, the main purpose of this work is to examine if individuals with musical training are more accurate in one area of voice-based social perception, i.e., age estimation. In this work, those individuals are defined to have "better skills" and be "more accurate" whose age judgments are closer to the calendar age of the speakers. As discussed later in detail, several authors pointed out that classical musical training enhances auditory processing skills in many ways. Does this lead to more accurate age judgments? If results support this, one can infer that voice-based social perception in musicians is based on more reliable foundations than in non-musicians, which may possibly influence decisions in social interactions.

1.2 Speaker age estimation

Humans are in general able to judge the age of the speaker, purely based on the voice, although certain inaccuracies exist. Among the first researchers, psychologists Allport and Cantril (1934) published results of age estimation. They found that estimates were centered around a median of 35-40 years of age, however, they only used three speakers (actual ages: 27, 36, 51 years). More advanced methodology with a larger sample of speakers was used by Shipp and Hollien (1969). Their results suggested that strong correlation exists between the actual and the perceived ages (r = 0.88). However, strong correlation does not imply accurate judgments. While listeners tended to underestimate older

speakers' age, they overestimated young adults' age (Huntley et al., 1987). Both the strong correlation between actual and estimated age and tendencies of underestimation and overestimation have been confirmed by many other researchers. For example, high degree of correlation between actual and estimated ages was found by Winkler et al. (2003) (r = 0.864 for spontaneous speech and 0.862 for reading), Cerrato et al. (2000), who used telephonic voices (r = 0.77), or Stölten and Engstrand (2002), who found r = 0.92 when their younger and the older speakers' groups were collapsed.

In speaker age estimation, two main approaches exist. Cross-sectional experiments use different speakers from a selected period, most commonly, recordings from speakers of the time when the research is carried out. In longitudinal settings, researchers use speech samples of the same persons recorded at different time points.

In a cross-sectional experiment, Hughes and Rhodes (2010) examined more closely how accurately the age of speakers belonging to different age groups can be judged. They found that raters were fairly accurate when determining the age of children and adolescents. A slight overestimation of speaker age was found with young male speakers, while the underestimation of the female speakers' age was more prominent for speakers in the age group of 23-34. The degree of underestimation, however, was higher for the male speakers in age groups 35-45 and 46-55. Finally, the age of speakers over 56 years of age was also underestimated but no significant difference between male and female speakers was found. Sandman et al. (2014) also reported a "tendency for estimates to gravitate to middle age", implying correct estimates for middle-aged speakers, while more prominent differences were found between the estimated and real ages for younger or older speakers.

A longitudinal experiment was carried out by Reubold et al. (2010). Two recordings from Queen Elizabeth II and three from broadcaster Alistair Cooke were used as stimuli. Listeners were accurate in identifying the Queen's age as younger when listening to a 1972 recording, while they were right to identify her as older when hearing a 1983 recording. Cooke was identified younger on a 1947 recording than on the 1970 recording; however, no significant differences between the 1970 and 1990 recordings were found. It was also found that the manipulation of f0, while other parameters remained constant, resulted in an unequivocal effect on perceptual age, while shifting F_1, while keeping f0 constant, provided mixed results. In another longitudinal study, 60 samples were extracted from public addresses by a male speaker over 48 years (aged 48-97). When the speaker was between 49 and 68 years old, he was estimated to be between 58 and 68, overestimating his age by 6 years on average. When the talker reached age 68, the estimates were in line with his calendar age; however, about 5 years of underestimation on average occurred (Hunter & Ferguson, 2017).

Concerning potential effects of listener factors, fewer results are available. Eppley and Mueller (2001) found that both the young and the old listener group underestimated the age of their elderly speakers; however, the older listeners' estimations were, on the average, some four years closer to the actual age of the speakers than those made by the younger group but the difference was not significant. These background factors were summarized by Moyse (2014), who concluded that younger listeners are more accurate than older ones, irrespective of the age of stimuli, and the age of female voices is more accurately estimated than that of male speakers.

Since the present study is carried out with Hungarian subjects, it is important to briefly review available literature. Previous findings with Hungarian speakers and listeners are in harmony with international results. A strong correlation between the real and estimated age (Gocsál, 1998; Bóna, 2013), inaccuracies of judgments, including overestimation or underestimation, have been documented in a variety of experimental settings (Gósy, 2001; Tatár, 2013; Tóth, 2014; Krepsz & Gósy, 2016; Gocsál, 2017).

1.3 Auditory perceptual skills of musicians

In the literature of speaker age estimation, listener factors such as age and gender have been considered so far. Little is known about other factors that may differentiate age judgments. One possible factor of this kind, which may play a role here, is the "quality" of the auditory processing mechanism of the listeners. How do listeners with more sophisticated auditory skills perceive extralinguistic contents of speech? Are they better at age estimation? One specific group of people in which auditory processing skills are expected to be better than those of others is musicians. For assessing potential differences in the auditory mechanisms of musicians and non-musicians, a large part of research uses non-linguistic acoustic stimuli. In most of the cases, musicians do demonstrate more sophisticated skills, i.e., they are more sensitive to smaller differences in the acoustic sign. For example, musicians are better at identifying changes of frequency of pure tones both in silent and noisy conditions (Liang et al., 2016). Further experimental evidence for enhanced performance on frequency discrimination (Micheyl et al., 2006; Eadie et al., 2010; Madikal Vasuki et al., 2016; Meha-Bettison et al., 2018), better auditory temporal-interval discrimination (Banai et al., 2012), or both (Boebinger et at,. 2015). Also, better pitch contour identification was found in musicians; however, training with pitch discrimination exercises results in the improvement of both musicians' and non-musicians' performance (Micheyl et al., 2006; Wayland et al., 2010).

In some cases, no such differences were found. For example, in tonal processing, sense of completion of a melody was rated similarly by musicians and non-musicians, even though neural responses were different. This suggests

that there may be specific mechanisms available for non-musicians, compensating for their lack of musical training (Amemiya et al., 2014).

Plasticity of the auditory cortex was also found to be induced by musical training, and Pantev and Herholz (2011) suggested that making music may even contribute to a more effective recovery from impaired auditory or motor skills caused by lesions. Musical training may also mitigate changes in auditory perception commonly occurring in aging adults (Alain et al., 2013).

Discrimination of pure tones does not necessarily explain better performance (Eadie et al., 2008), spectrally more complex stimuli may help determine the underlying differences. It is therefore of particular interest to examine language related perceptual skills. In this context, effects of music training were studied by Flaugnacco et al. (2015). Their findings suggest that music training boosts phonological awareness, rhythmic abilities and reading skills, even when these skills are impaired. Speech segmentation skills are also improved by music training, which may contribute to children's language development (François et al., 2013). Musicians are also more sensitive to subtle pitch changes, both in non-linguistic tones and spoken sentences (Deguchi et al., 2012). Kühnis et al. (2013) used spectrally and temporally manipulated CV syllables, and musicians demonstrated an increased responsiveness to the acoustic stimuli. Alexander et al. (2005) found that English-speaking musicians performed significantly better at identifying and discriminating Chinese lexical tones than English-speaking non-musicians.

Musicianship seems also to be an advantage when one perceives speech in noise. Better perceptual abilities of musicians were demonstrated by Parbery-Clark et al. (2009, 2011). Positive effects of music training on speech in noise perception were also demonstrated in children by a longitudinal research (Stater et al., 2015). A review paper by Coffey et al. (2017) compared research results obtained over a wide range of conditions and confirmed musicians' advantage in speech-in-noise perception. However, some results do not support this conclusion: for example, those found by Boebinger et al. (2015) who concluded that it was nonverbal IQ rather than musical training that predicted speech perception thresholds in noise. A more realistic version of this type of experiment uses the multi-talker masking approach where the masking noise simulates a "cocktail-party" environment (Swaminathan et al., 2015). Musicians outperformed non-musicians when the maskers were spatially separated from the target voice, but no significant differences were found when the masking voices and the target voices were collocated.

In another experiment by Sadakata and Sekiyama (2011), musicians were also better at discriminating and identifying morphed speech sounds, that is, they were more sensitive to subtle temporal and timbre differences of speech sounds when they heard minimal pairs of words, one of them unaltered, while the other

chosen from a series of slightly morphed versions of the words. Deme's (2017) results, however, suggest that professional singers do not enjoy a perceptual advantage in the identification of high-pitched, sung vowels over naïve listeners.

Concerning speech and music perception, only few comparative studies have been published. Chartrand and Belin (2006) found that in discriminating instrumental sounds and human voice samples with different timbres, musicians outperformed non-musicians. Musicians used more response time though, which was most likely due to a deeper level processing of the sounds. Another question related to the link between musicianship and speech perception addresses the issue of disorders related to music perception. An experiment by Liu et al. (2015) demonstrated that individuals who suffer from congenital amusia, i.e. are unable to discriminate pitch levels of a melody from birth, also experience difficulties in speech comprehension. Their results also revealed that amusia is not limited only to pitch processing. Amusic subjects achieved lower scores in perceiving flat f0 sentences, which implies that their deficits in speech perception go beyond pitch processing. These results are worth noting because they contradict statements that amusic individuals have a normal understanding of speech (Peretz & Vuvan, 2017). A wide range of other research related to differences in musicians' auditory skills is available in the literature. For example, when evaluating dysphonic voices, musicians demonstrated significantly more agreement in judging the breathiness in dysphonic speakers than non-musicians (Eadie et al., 2008).

The question whether auditory processing mechanisms for voice and music are separate or are at least in part overlapping has been discussed by several researchers. Chartrand and Belin (2006) proposed that voice timbre and instrument timbre discrimination involve similar mechanisms. If music related perceptual mechanisms were completely independent from voice related mechanisms, experiments would only demonstrate musicians' advantage of discriminating instrument timbres, but not vocal timbres. However, their results showed that musicians outperformed non-musicians both in voice and instrument timbre discrimination tasks as well, thus, overlapping perceptual mechanisms were suggested. Hausen et al. (2013) found that music and speech perception is shared by the perception of rhythm and pitch. Strait and Kraus (2011) mentioned several specific common mechanisms such as auditory attention, working memory, neural function in challenging listening environments, sequential sound processing, and sensitivity to temporal and spectral aspects of sound. Perhaps the most comprehensive model that connects the perception of speech and music is proposed by Patel (2014). His expanded OPERA hypothesis (O = overlap between neural networks processing speech and music, P = higher precision of processing is demanded when music is heard, E = emotion, R = repetition of musical activities, A = focused attention

demanded by music), proposes that when music and speech share auditory perceptual mechanisms, and music places higher demands on those auditory mechanisms than speech does, speech processing may be enhanced in musicians (Patel, 2014). One possible explanation is that musical training improves the generic constitutional properties of the auditory system (Kühnis et al., 2013).

In the context of the present research, perception of non-linguistic contents of speech is in focus. One direction of this kind of research is related to the perception of emotions. More intense patterns of emotional activation were observed in musicians than in non-musicians when they were listening to classical music (Mikutta et al., 2014), and out of the different types of emotive stimuli, negative emotions expressed in music are more arousing for musicians while responses to happiness in music evoke no specific activations (Park et al., 2014). It is therefore an interesting question to examine how musicians perceive emotions from speech. Results suggest that musicians have better skills in identifying the emotions conveyed by tone sentences mimicking the prosody of spoken sentences. Also, musically trained adults were better at identifying spoken utterances that were emotionally neutral (Thompson et al., 2004). In an experiment by Pinheiro et al. (2015), musicians were more accurate in recognizing angry prosody from sentences, and, their general conclusion was that extensive musical training may impact different stages of vocal emotional processing.

1.4 Research questions

It has been demonstrated that humans are capable of estimating the speaker's age, based on the speaker's voice, although certain inaccuracies occur in estimations. Previous studies have found that speech rate, and in certain cases, speaking fundamental frequency are key properties of speech that listeners use (Reubold et al., 2010; Stölten & Engstrand, 2003; Winkler, 2007; Skoog Waller et al., 2012) in age judgments. If musicians' auditory skills are more sensitive to temporal and frequency differences (Tervianemi et al., 2005; Elmer et al., 2012), it seems reasonable to raise the question if musicians' age judgments are closer to the speakers' calendar age than those of non-musicians. In particular, the research questions are as follows: (1) Do musicians' and non-musicians' age estimations correlate with the speakers' calendar age? (2) Do musicians' age estimations differ from those of non-musicians in different age groups of speakers? (3) Do differences in male and female listeners exist? Since we assume that musicians have more sophisticated auditory mechanisms, therefore it is hypothesized that (1) musician listeners' age estimations correlate stronger with the speakers' calendar ages than non-musicians' estimations and (2) in each of three different age groups of speakers, musicians' age estimations are more

accurate than those of non-musicians, and (3) we expect no differences between male and female listeners' age estimations.

2 Methods

2.1 Acoustic stimuli

24 male speakers (age range: 20-72 years) were selected from the BEA spontaneous speech database (Gósy et al., 2012) in a way that their ages were approximately evenly distributed over the age range of the whole sample, resulting a mean difference of 2.16 years (range: 0-4 years) between two adjacent speakers. The speakers were nonsmokers and had either BSc or higher degrees or were students. Of the recordings, samples of approx. 20-30 seconds were chosen from the "interview" or "argument" task of the BEA recording protocol. The chosen speech samples were from longer monologues in which the speakers were talking about general themes, e.g. their job, local transportation, hobbies etc. in an emotionally neutral way. The speech samples included no textual information that the listeners may have used for inferring the speakers' age.

2.2 Listeners

Listeners were students of the University of Pécs (n = 85, age range: 19-37, median: 22), without any prior training in phonetics. There were 42 listeners who study music (instrumental players of classical music, with at least 8 years of music education), and 43 students of other fields (sociology, fine arts, media), who cannot play an instrument and never received any kind of music training, apart from the compulsory singing classes at school. There were some non-musician participants who had received some music training, but they were excluded from the study. Table 1 shows the gender distribution of musician and non-musician participants. No participant reported any kind of hearing impairment, complaints or previous medical treatment that may have influenced their auditory processes.

Table 1. Participants of the study

	musicians	**non-musicians**
males	14	14
females	28	29

2.4 Procedure

The listening tasks took place in a silent seminar room of the Zsolnay campus of the University of Pécs, through good quality multimedia speakers, in groups of 5-10. Listeners were instructed to estimate the speaker's age in years and to write down their estimations on an answer sheet. Prior to the experiment, listeners were familiarized with the task by playing three sound samples to them. At the same time, the experimenter tested if all the listeners can clearly hear the

sound samples in the seminar room. Each of the sound samples was introduced by a 1 second 440 Hz beeping sound and a 2 second silent pauses. Each sound sample was played once. A sound sample was only played when all the subjects were ready with writing down their estimations for the previous sample. The speech samples were played in the same, randomized order in each group. Once the data were obtained, Pearson's correlation coefficients were calculated and linear mixed models were fitted using the SPSS 23.0 software.

3 Results

3.1 Correlations between mean estimated age and real age

To establish possible similarities with previous findings, mean values of the age estimations were calculated for each speaker, and Pearson's correlation coefficients between the mean estimated ages and calendar ages were calculated. Figure 1 shows the scatterplot of the estimated mean ages against calendar ages with all listener groups collapsed. Pearson's correlation coefficient indicates strong correlation ($r = 0.806$, $p < 0.001$). A strong correlation, however, does not necessarily suggest accurate judgments. The deviation of the regression line (dashed line) from the y = x line (solid line) suggests a tendency for younger speakers to be perceived older, while older speakers are believed to be younger than their actual age. The intersection of the two lines suggests accurate age estimations for speakers between 35 and 40 years of age.

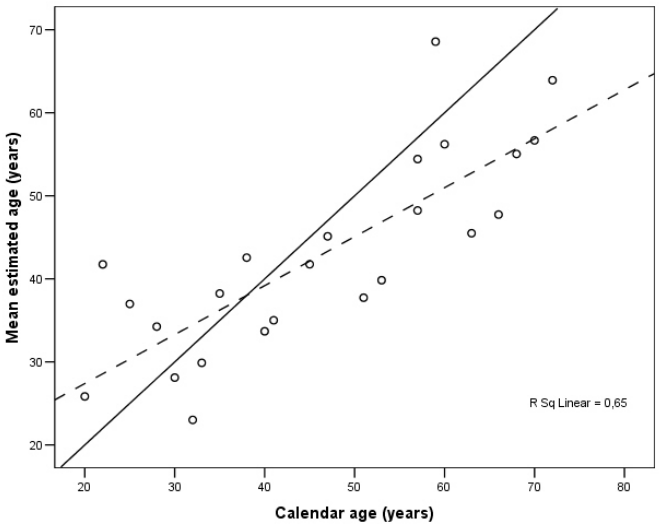


Figure 1.
Scatterplot of mean estimated age and calendar age,
all listener groups collapsed

Figure 2 shows musician participants' mean age estimations against the calendar age of the speakers. For both musician males and females, correlation coefficient is comparable or identical with what was found with the overall group ($r = 0.803$ in males and $r = 0.806$ in females, $p < 0.001$ in both cases). The regression lines predict virtually identical age estimations in younger speakers, while female listeners' age estimations are predicted to be slightly closer to the real age of the older speakers.

Figure 3 illustrates non-musician participants' mean age estimations against the calendar age of the speakers.

Again, significant correlation coefficients are found (non-musician males: $r = 0.839$, non-musician females: $r = 0.777$, $p < 0.001$), which is slightly stronger in males and slightly weaker in females than what was found with the musicians' group. The two regression lines are almost identical.

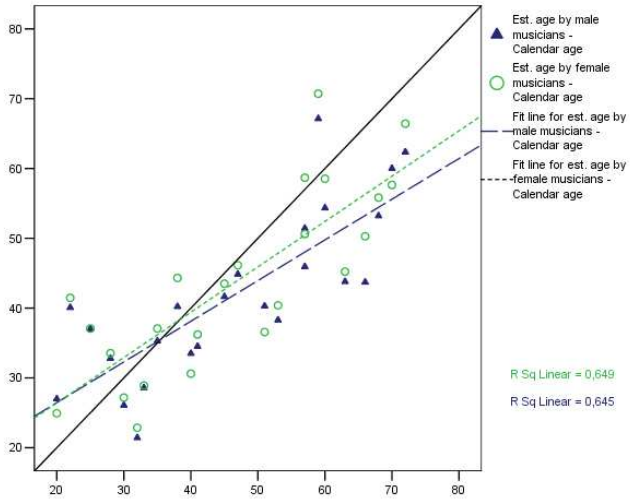


Figure 2.
Scatterplot of mean estimated age and calendar age, musician listeners

Next, to explore whether listener gender and musicianship significantly influence age estimations, a linear mixed model was fitted with estimated age as dependent variable, calendar age as covariate and listener gender and musicianship as fixed factors. For this analysis, all data were used rather than mean values of age estimations. The obtained F-values were as follows: $F(1, 70.591) = 0.988$, $p = 0.324$ for listener gender, $F(1, 70.591) = 0.562$, $p = 0.456$ for musicianship and $F(1, 70.591) = 0.202$, $p = 0.654$. These values suggest that neither the individual fixed factors, nor their interaction is significant. In sum, although minor differences were observed between the listener groups when

Pearson's correlation coefficients were used, a deeper analysis, including all data rather than mean values, revealed no effect of listener gender and musicianship.

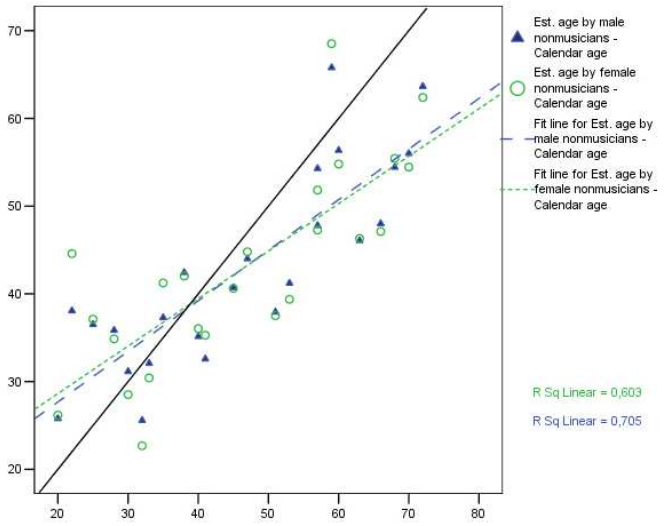


Figure 3.
Scatterplot of mean estimated age and calendar age, non-musician listeners

3.2 Accuracy of age estimations in different speaker age groups

To determine the nature of accuracy of speaker age estimations, linear mixed models were implemented again in a different way. First, for each speaker, differences between the estimated and the calendar ages were calculated, and the difference values were z-standardized. Of the resulting 2040 values (85 listeners x 24 speakers) seven (0.3%) were excluded because of the z-score being over 3 or below –3. Figure 4 demonstrates the related boxplots for all speakers.

Although a visual inspection of the boxplots would suggest a slight difference in the mean values, calculations do not confirm a significant difference. Several covariate structures were tested for repeated effects, but none resulted in significant fixed effects. The lowest AIC value was obtained with the covariate structure "scaled identity". The resulting F-values are as follows: $F(1, 81.058) = 0.089$, $p = 0.766$ for the intercept, $F(1, 81.058) = 1.128$, $p = 0.291$ for the gender of the listener, $F(1, 81.058) = 0.341$, $p = 0.570$ for musicianship, and $F(1, 81.058) = 1.126$, $p = 0.292$ for the interaction of gender and musicianship. These results suggest that there is no significant tendency of underestimation or overestimation of age between the listener groups, i.e., it cannot be stated that when all speakers are collapsed into a single group, age estimations of musicians and non-musicians, males and females significantly differ.

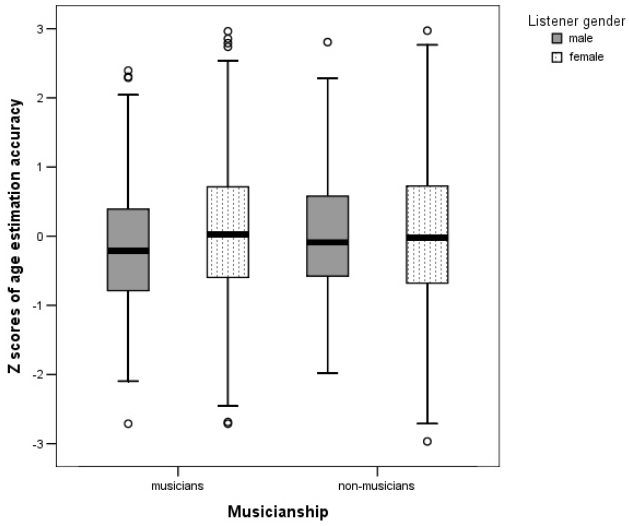


Figure 4.
Boxplot of *z*-scores of age estimations for all speakers

It should, however, be noted that the above calculation does not rule out possible differences in different age groups. For a further analysis, three groups were made of the speakers and the same calculations were repeated separately with the groups. The young speakers' group included speakers where overestimation of age was most likely. The middle-aged speakers' group, accurate estimations were expected, and, in the older speakers' group, underestimation of age was most likely. Again, in each group, *z*-scores over 3 or below –3 were excluded. First, age estimations for 5 speakers, between 22 and 30, were analyzed. Figure 5 shows the results.

Again, a visual inspection of the boxplots shows lower *z*-scores for the male musicians. Since the age of the younger speakers was in general overestimated, lower *z* values imply smaller differences between the estimated age and the calendar age, i.e. better age estimations. In the linear mixed model, different covariate structures resulted in very similar outcomes. Applying the covariate structure "scaled identity", the following *F*-values were obtained: $F(1, 85.339) = 0.119$, $p = 0.731$ for the intercept, $F(1, 85.339) = 0.232$, $p = 0.631$ for the gender of the listener, $F(1, 85.339) = 3.411$, $p = 0.068$ for musicianship, and $F(1, 85.339) = 0.000$, $p = 1.000$ for the interaction of gender and musicianship. These findings suggest a marginal but not significant main effect of musicianship, i.e., the lower z-scores of musician males suggest slightly more accurate estimations than those of the other groups, while non-musician females

seem to have been less accurate than the others. But, again, this difference is not significant.

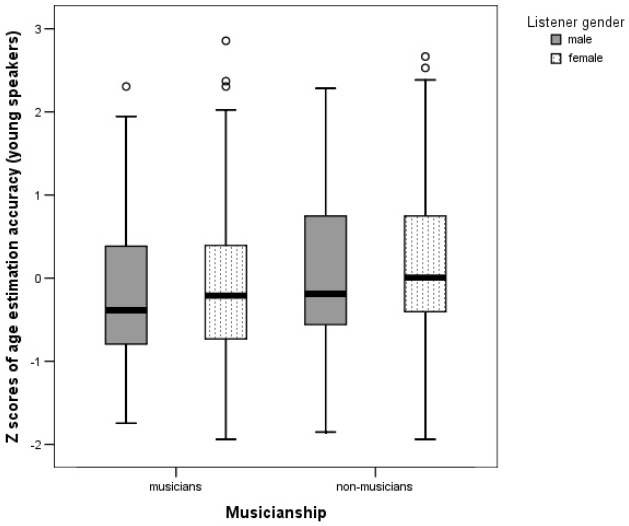


Figure 5.
Boxplot of *z*-scores of age estimations for young speakers

Finally, the calculations were carried out with the older speakers' group (10 speakers, ages between 53-72 years). In this case, one listener group, musician females, seems to be prominent, as Figure 7 shows.

The second group consisted of 9 speakers between 32 and 50 years of age. Figure 6 shows slight differences between the boxplots; however, the differences are not significant. Again, different covariate structures result in substantially the same outcomes. The results, i.e., $F(1, 84.804) = 0.036$, $p = 0.850$ for the intercept, $F(1, 84.804) = 0.510$, $p = 0.477$ for the gender of the listener, $F(1, 84.804) = 1.255$, $p = 0.266$ for musicianship, and $F(1, 84.804) = 0.130$, $p = 0.723$ for the interaction of gender and musicianship, demonstrate not even a marginal effect of any of the factors.

Again, several covariate structures were tested but substantially the same results were obtained. With the "scaled identity" covariate structure, intercept ($F(1, 85.073) = 0.057$, $p = 0.812$) and musicianship ($F(1, 85.073) = 0.709$, $p = 0.402$) were not significant fixed factors, neither was gender ($F(1, 85.073) = 1.349$, $p = 0.249$) or the gender × musicianship interaction ($F(1, 85.073) = 3.047$, $p = 0.085$). Removal of intercept did not improve the effect of the gender × musicianship interaction. This marginal but not significant interaction suggests somewhat better age estimations of female musicians (see Figure 7). In this case,

higher *z*-values imply better age estimations since higher estimated ages were closer to the calendar age.

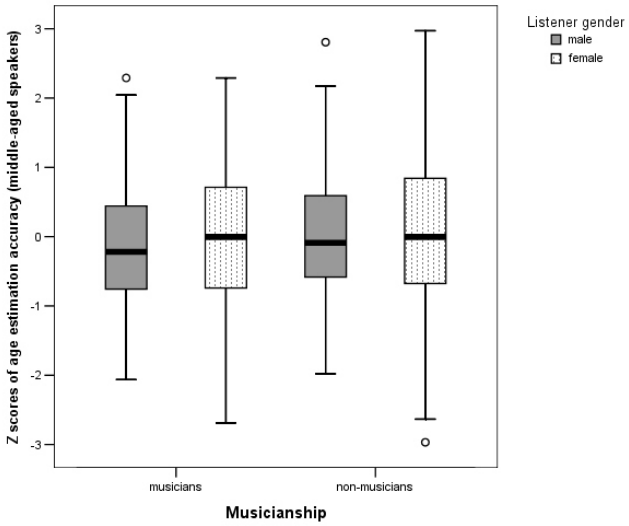


Figure 6.
Boxplot of *z*-scores of age estimations for middle-aged speakers

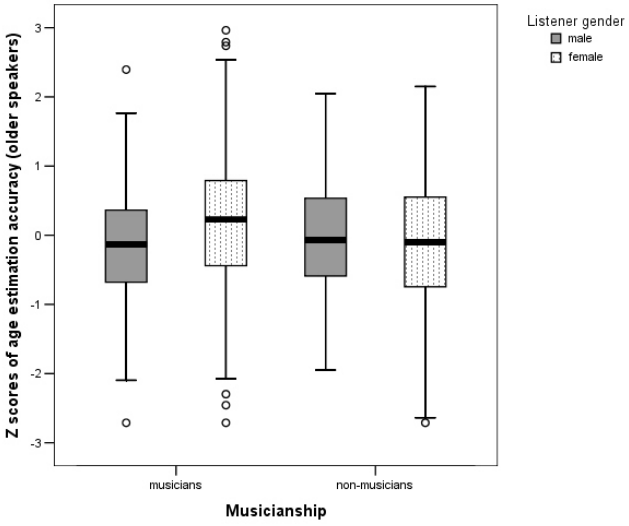


Figure 7.
Boxplot of *z*-scores of age estimations for older speakers

4 General discussion

The main purpose of the present paper was to reveal if musicians are more accurate in speaker age estimation than non-musicians. Our first hypothesis was not confirmed. Pearson correlation coefficients between mean estimated ages and calendar ages (*r*-values between 0.777 and 0.838) are in line with those reported in previous studies (Huntley et al., 1987; Cerrato et al., 2000; Winkler et al., 2003), and differ between the listener groups, but a repeated measures linear mixed model revealed no significant effect of listener gender and musicianship.

Our second hypothesis was not confirmed either. No statistically significant differences were demonstrated by further analyses carried out with different age groups. The third hypothesis was confirmed, no significant differences were found between male and female listeners' estimations. However, there was a non-significant tendency for male musicians to be more accurate in the age estimation of younger speakers, while female musicians were marginally better at estimating the age of the older speakers.

One possible explanation for not finding more prominent differences is that differences between the vocal parameters of different aged speakers are large enough to be perceived even by listeners with less sophisticated perceptual mechanisms. The marginally better performance of male and female musicians in the two cases may reflect that they have "more accurate" vocal prototypes, at least for those age groups. Krepsz and Gósy (2016) outlined the nature of such vocal prototypes, which one builds through experiencing speakers and voices. Vocal prototypes then serve as bases for social perception, including age estimation. Future research should address this marginal difference. Other research has pointed out the role of voice quality in assessing the speaker's dominance or attractiveness (Fraccaro et al., 2013; Puts et al., 2014) so it is possible that for male listeners it is important to develop vocal stereotypes so that they are more accurate in the perception of peers, while, for female listeners, it may be important to develop skills to be more accurate in the perception of more mature males. Musical training may be an advantage in the development of such vocal stereotypes and skills, but again, since non-significant differences were found, further experiments are needed.

One of the main limitations of the present study is that the role of acoustic parameters was not analyzed. Previous studies have demonstrated the role of speech rate (Skoog Waller et al., 2015; Gocsál, 2017), fundamental frequency and formant structure (Reubold et al., 2010), or a combination of several parameters (Winkler, 2007; Harnsberger et al., 2008) in age estimations. Future work should also address the question if musicians and non-musicians use different strategies in using the acoustic parameters of speech in speaker age estimation. Their marginally better estimations in some cases may be attributable

to their attention that they pay to subtle details in the acoustic structure of speech. However, the effect of those subtle details is not large enough to make significantly better estimates. Also, further studies will need to examine why female listeners' estimations covered a larger range than those of males (especially with speakers in the middle-aged and older group, see Figures 6 and 7). It is possible that they are more uncertain in giving estimations on opposite-sex speakers, but a similar experiment with female speakers would be necessary for a comparison.

In conclusion, it can be stated that this experiment was the first attempt to demonstrate potential benefits of musicianship in speaker age estimation, and although no prominent differences were found, the marginally better estimations of musicians in some cases raise questions that deserve attention in the future.

References

Abend, P., Pflüger, L. S., Koppensteiner, M., Coquelle, M., & Grammer, K. (2015). The Sound of Female Shape: A Redundant Signal of Vocal and Facial Attractiveness. *Evolution and Human Behavior, 36,* 174-181.

Alain, C., Zendel, B. R., Hutka, S., & Bidelman, G. M. (2013). Turning down the noise: The benefit of musical training on the aging auditory brain. *Hearing Research, 308*, 162-173.

Alexander, J. A., Wang, P. C. M., & Bradlow, A. R. (2005). Lexical tone perception in musicians and non-musicians. In *9th European Conference on Speech Communication and Technology, Eurospeech Interspeech* (pp. 397–400). Lisbon, Portugal.

Allport, G. W., & Cantril, H. (1934). Judging personality from voice. *Journal of Social Psychology: Political, Racial and Differential Psychology, 5*, 37-55.

Amemiya, K., Karino S., Ishizu, T., Yumoto, M., & Yamasoba, T. (2014). Distinct neural mechanisms of tonal processing between musicians and non-musicians. *Clinical Neurophysiology, 125,* 738-747.

Banai, K., Fisher, S, & Ganot, R. (2012). The effects of context and musical training on auditory temporal-interval discrimination. *Hearing Research, 284,* 59-66.

Boebinger, D., Evans, S., Scott, S. K., Rosen, S., Lima, C. F., & Manly, T. (2015). Musicians and non-musicians are equally adept at perceiving masked speech. *Journal of the Acoustic Society of America, 137*(1), 378-387.

Bóna, J. (2013). A spontán beszéd sajátosságai az időskorban [Spontaneous speech in old age]. Budapest: ELTE Eötvös Kiadó.

Cargile, A. C., & Giles, H. (1997). Understanding language attitudes: Exploring listener affect and identity. *Language & Communication, 17*(3). 195-217.

Cerrato, L., Falcone, M., & Paolini, A. (2000). Subjective age estimation of telephonic voices. *Speech Communication, 31*, 107-112.

Chartrand, J. P., & Belin, P. (2006). Superior voice timbre processing in musicians. *Neuroscience Letters, 405*, 164-167.

Charlton, B. D., Taylor A.M., & Reby, D. (2013). Are men better than women at acoustic size judgements? *Biology Letters, 9,* 1-5.

Coffey, E. B. J., Mogilever, N. B., & Zatorre, R. J. (2017). Speech-in-noise perception in musicians: A review. *Hearing Research, 352*, 49-69.

Collins, S. (2000). Men's voices and women's choices. *Animal Behaviour* 60. 773–780.

Cross, J. B., DeVaney, T., & Jones, G. (2001). Pre-service teacher attitudes toward differing dialects. *Linguistics and Education, 12*(4), 211-227.

Deguchi, Ch., Boureux, M., Sarlo, M., Besson, M., Grassi, M., Schön, D., & Colombo, L. (2012). Sentence pitch change detection in the native and unfamiliar language in musicians and non-musicians: Behavioral, electrophysiological and psychoacoustic study. *Brain Research, 1455*, 75-89.

Deme, A. (2017). The identification of high-pitched sung vowels in sense and nonsense words by professional singers and untrained listeners. *Journal of Voice, 31*(2), 252.e1-252.e14.

van Dommelen, W. A., & Moxness, Bente H. (1995). Acoustic parameters in speaker height and weight identification: Sex-specific behavior. *Language and Speech, 38*(3), 267-287.

Drager, K. 2010. Sociophonetic variation in speech perception. *Language and Linguistics Compass, 4*(7), 473-480.

Eadie, T., Van Boven, L., Stubbs, K., & Giannini, E. (2010). The effect of musical background on judgments of dysphonia. *Journal of Voice, 24*(1). 93-101.

Elmer, S., Meyer, M., & Jäncke, L. (2012). Neurofunctional and behavioral correlates of phonetic and temporal categorization in musically trained and untrained subjects. *Cerebral Cortex, 22,* 650-658.

Eppley, B. D., & Mueller, P. B. (2001). Chronological age judgments of elderly speakers: The effects of listeners' age. *Contemporary Issues in Communication Science and Disorders, 28,* 5-8.

Feinberg, D. R., Jones, B. C., Little, A. C., Burt, M. D., & Perrett, D. I. (2005). Manipulations of fundamental and formant frequencies influence the attractiveness of human male voices. *Animal Behaviour, 69,* 561-568.

Flaugnacco, E., Lopez, L., Terribili, C., Montico, M., Zoia, S., & Schön, D. (2015). Music training increases phonological awareness and reading skills in developmental dyslexia: A randomized control trial. *PLoS ONE, 10*(9). e0138715. doi:10.1371/journal.pone.0138715

Fraccaro, P. J., O'Connor, J. J. M., Re, D. E., Jones, B. C., DeBruine, L. M., & Feinberg, D. R. (2013). Faking it: deliberately altered voice pitch and vocal attractiveness. *Animal Behaviour, 85,* 127-136.

François, C., Chobert, J., Besson, M., & Schön, D. (2013). Music training for the development of speech segmentation. *Cerebral Cortex, 23,* 2038-2043. doi:10.1093/cercor/bhs180

Gelfer, M. P., & Bennett, Q. E. (2014). Speaking fundamental frequency and vowel formant frequencies: effects on perception of gender. *Journal of Voice, 27*(5), 556-566.

Gocsál, Á. (1998). Életkorbecslés a beszélő hangja alapján [Age estimation based on the speaker's voice]. *Beszédkutatás '98,* 122-134.

Gocsál, Á. (2017). Az artikulációs tempó és az átlagos alaphang szerepe a beszélő életkorának megbecslésében [Age estimation based on the mean f0]. *Beszédkutatás 2017,* 151-168.

Gósy, M. 2001. A testalkat és az életkor becslése a beszéd alapján [Estimation of body configuration and age based on speech]. *Magyar Nyelvőr, 125,* 137-148.

Gósy, M., Gyarmathy, D., Horváth V., Gráczi. T. E., Beke, A., Neuberger, T., & Nikléczy, P. 2012. BEA: beszélt nyelvi adatbázis [BEA: Spoken language database]. In M. Gósy (Ed.), *Beszéd, adatbázis, kutatások* [Speech, darabase, research] (pp. 9-24). Budapest: Akadémiai Kiadó.

Harnsberger, J. D., Shrivastav, R., Brown Jr., W. S., Rothman, H., & Hollien, H. (2008). Speaking rate and fundamental frequency as speech cues to perceived age. *Journal of Voice, 22*(1), 58-69.

Hausen, M., Torppa, R., Salmela, V. R., Vainop, M., & Särkämö, T. (2013). Music and speech prosody: A common rhythm. *Frontiers in Psychology, 4*(566). 1-16. doi: 10.3389/fpsyg.2013.00566

Hughes, S. M., & Rhodes, B. C. (2010). Making age assessments based on voice: The impact of the reproductive viability of the speaker. *Journal of Social, Evolutionary, and Cultural Psychology*, *4*(4), 290-304.

Hughes, S. M., & Harrison, M. A. (2013). I like my voice better: Self-enhancement bias in perceptions of voice attractiveness. *Perception, 42,* 941-949.

Hunter, E. J., & Ferguson, S. H. (2017). Listener estimates of talker age in a single-talker, 50-yearlongitudinal sample. *Journal of Communication Disorders, 68,* 103-112.

Huntley, R., Hollien, H. & Shipp, T. (1987). Influences of listener characteristics on perceived age estimations. *Journal of Voice,* 1(1), 49-52.

Klofstad, C. A., Anderson, R. C., & Nowicki, S. (2015). Perceptions of competence, strength, and age influence voters to select leaders with lower-pitched voices. *PLoS ONE* 10(8): e0133779.doi:10.1371/journal.pone.0133779 (Retrieved 15.03.2018).

Krauss, R., Freyberg, R., & Morsella, E. (2002). Inferring speakers' physical attributes from their voices. *Journal of Experimental Social Psychology, 38*, 618-625.

Krepsz, V., & Gósy, M. (2016). A hangzásidő és a megakadásjelenségek hatása az életkorbecslésre [The effect of the duration of sounding and the disfluencies on the estimation of the speaker's age.]. In G. Balázs, & Á. Veszelszki (Eds.), *Generációk nyelve. Tanulmánykötet.* [Languages of Generations. Book of selected papers] (49–62). ELTE BTK Mai Magyar Nyelvi Tanszék, Inter Nonnprofit Kft. – MSZT, Budapest.

Kühnis, J., Elmer, S., Meyer, M., & Jäncke, L. (2013). The encoding of vowels and temporal speech cues in the auditory cortex of professional musicians: An EEG study. *Neuropsychologia, 51,* 1608-1618.

Liang, C., Earl, B., Thompson, I., Whitaker, K., Cahn, S. Xiang, J., Fu, Q.-J., & Zhang, F. (2016). Musicians are better than non-musicians in frequency change detection: Behavioral and Electrophysiological evidence. *Frontiers in Neuroscience, 10,* 464. doi: 10.3389/fnins.2016.00464

Liu F., Jiang, C., Wang, B., Xu, Y., & Patel, A. D. (2015). A music perception disorder (congenital amusia) influences speech comprehension. *Neuropsychologia, 66,* 111-118.

Madikal Vasuki, P. R., Sharma, M., & Demuth, J. A. (2016). Musicians' edge: A comparison of auditory processing, cognitive abilities and statistical learning. *Hearing Research, 342,* 112-123.

Meha-Bettison, K., Sharma, M., Ibrahim, R. K., & Vasuki P. R. M. (2018). Enhanced speech perception in noise and cortical auditory evoked potentials in professional musicians. *International Journal of Audiology, 57*(1), 40-52.

Micheyl, Ch., Delhommeau, K., Perrot, X. & Oxenham, A. J. (2006). Influence of musical and psychoacoustical training on pitch discrimination. *Hearing Research, 219,* 36-47.

Mikutta, C. A., Maissen, G., Altorfer, A., Strik, W., & Koenig, T. (2014). Professional musicians listen differently to music. *Neuroscience, 268,* 102-111.

Moyse, E. (2014). Age Estimation from Faces and Voices: A Review. *Psychologica Belgica, 54*(3), 255-265.

Pabery-Clark, A., Skoe, E., & Kraus, N. (2009). Musical experience limits the degradative effects of background noise on the neural processing of sound. *The Journal of Neuroscience, 29*(45), 14100-14107.

Pabery-Clark, A., Strait, D. L., & Kraus, N. (2011). Context-dependent encoding in the auditory brainstem subserves enhanced speech-in-noise perception in musicians. *Neuropsychologia, 49,* 3338-3345.

Pantev, C., & Herholz, S. C. (2011). Plasticity of the human auditory cortex related to musical training. *Neuroscience and Biobehavioral Reviews, 35,* 2140-2154.

Park, M., Gutyrchik, E., Bao, Y., Zaytseva, Y., Carl, P., Welker, L., Pöppel, E., Reiser, M., Blautzik, J., & Meindl, T. (2014). Differences between musicians and non-musicians in neuro-affective processing of sadness and fear expressed in music. *Neuroscience Letters, 566,* 120-124.

Patel, A. D. (2014). Can nonlinguistic musical training change the way the brain processes speech? The expanded OPERA hypothesis. *Hearing Research, 308,* 98-108.

Peretz, I., & Vuvan, D. T. (2017). Prevalence of congenital amusia. *European Journal of Human Genetics 2017,* 1-6.

Pinheiro, A. P., Vasconcelos, M., Dias, M., Arrais, N., & Gonçalves, Ó. F. (2015). The music of language: An ERP investigation of the effects of musical training on emotional prosody processing. *Brain & Language, 140,* 24-34.

Pisanski, K., Oleszkiewicz, A., & Sorokowska, A. (2016). Can blind persons accurately assess body size from the voice? *Biology Letters, 12,* 20160063.

Puts, D. A., Gaulin, S. J. C., & Verdolini, K. (2006). Dominance and the evolution of sexual dimorphism in human voice pitch. *Evolution and Human Behavior,* 27, 283-296.

Puts, D. A., Doll, L. M., & Hill A. K. (2014). Sexual Selection on Human Voices. In V. Weekes-Shackelford, & T. Shackelford (Eds.), *Evolutionary Perspectives on Human Sexual Psychology and Behavior. Evolutionary Psychology* (pp. 69-86). New York, NY: Springer.

Rendall, D., Vokey, J. R., & Nemeth, C. (2007). Lifting the curtain on the wizard of Oz: Biased voice-based impressions of speaker size. *Journal of Experimental Psychology: Human Perception and Performance, 33*(5), 1208-1219.

Reubold, U., Harrington, J., & Kleber, F. (2010). Vocal aging effects on f0 and the first formant: A longitudinal analysis in adult speakers. *Speech Communication, 52,* 638-651.

Rubin, D., DeHart, J., & Heintzman, M. (1991). Effects of accented speech and culture-typical compliance-gaining style on subordinates' impressions of managers. *International Journal of Intercultural Relations, 15,* 267-283.

Sadakata, M., & Sekiyama, K. (2011). Enhanced perception of various linguistic features by musicians: A cross-linguistic study. *Acta Psychologica, 138,* 1-10.

Sandman, K., am Zehnhoff-Dinnesen, A., Schmidt, C-M., Rosslau, K., Lang-Roth, R., Burgmer, M., Knief, A., Matulat, P., Vauth, M., & Deuster, D. (2014). Differences between self-assessment and external rating of voice with regard to sex characteristics, age, and attractiveness. *Journal of Voice, 28*(1), 128. e11-128.e18.

Shipp, Th., & Hollien, H. (1969). Perception of the aging male voice. *Journal of Speech, Language, and Hearing Research, 12,* 703-710.

Shoup-Knox, M. L., & Pipitone, R. N. (2015). Physiological changes in response to hearing female voices recorded at high fertility. *Physiology & Behavior, 139,* 386-392.

Skoog Waller, S., Eriksson, M., & Sörqvist, P. (2015). Can you hear my age? Influences of speech rate and speech spontaneity on estimation of speaker age. *Frontiers in Psychology, 6*(978). 1-11.

Stölten, K., & Engstrand, O. (2002). Effects of sex and age in the Arjeplog dialect: a listening test and measurements of preaspiration and VOT. *Proceedings of Fonetik, TMH-QPSR, 44*(1). 029-032. Online: http://www.speech.kth.se/prod/publications/files/qpsr/2002/2002_44_1_029-032.pdf (retrieved 07.12.2017)

Stölten, K., & Engstrand, O. (2003). Effects of perceived age on perceived dialect strength: A listening test using manipulations of speaking rate and f0. *PHONUM, 9,* 29-32.

Strait, D., & Kraus, N. (2011). Playing music for a smarter ear: cognitive, perceptual and neurobiological evidence. *Music Perception, 29*(2), 133-146.

Swaminathan, J., Mason, C. R., Streeter, T. M., Best, V., Kidd, G., & Patel, A. D. (2015). Musical training, individual, differences and the cocktail party problem. *Scientific Reports,* 5. Article number: 11628. doi:10.1038/srep11628

Tatár, Z. (2013). Beszélőprofil-alkotás lehetőségei a kriminalisztikai fonetikában [Possibilities of speaker profiling in forensic phonetics]. *Alkalmazott Nyelvtudomány, XIII*(1-2), 121-130.

Tervianemi, M., Just, V., Koelsch, S., Widmann, A., & Schröger, E. (2005). Pitch discrimination accuracy in musicians vs nonmusicians: an event-related potential and behavioral study. *Experimental Brain Research, 161*(1), 1-10.

Thompson W. F., Scellenberg, E. G., & Husain, G. (2004). Decoding speech prosody: do music lessons help? *Emotion, 4*(1), 46-64.

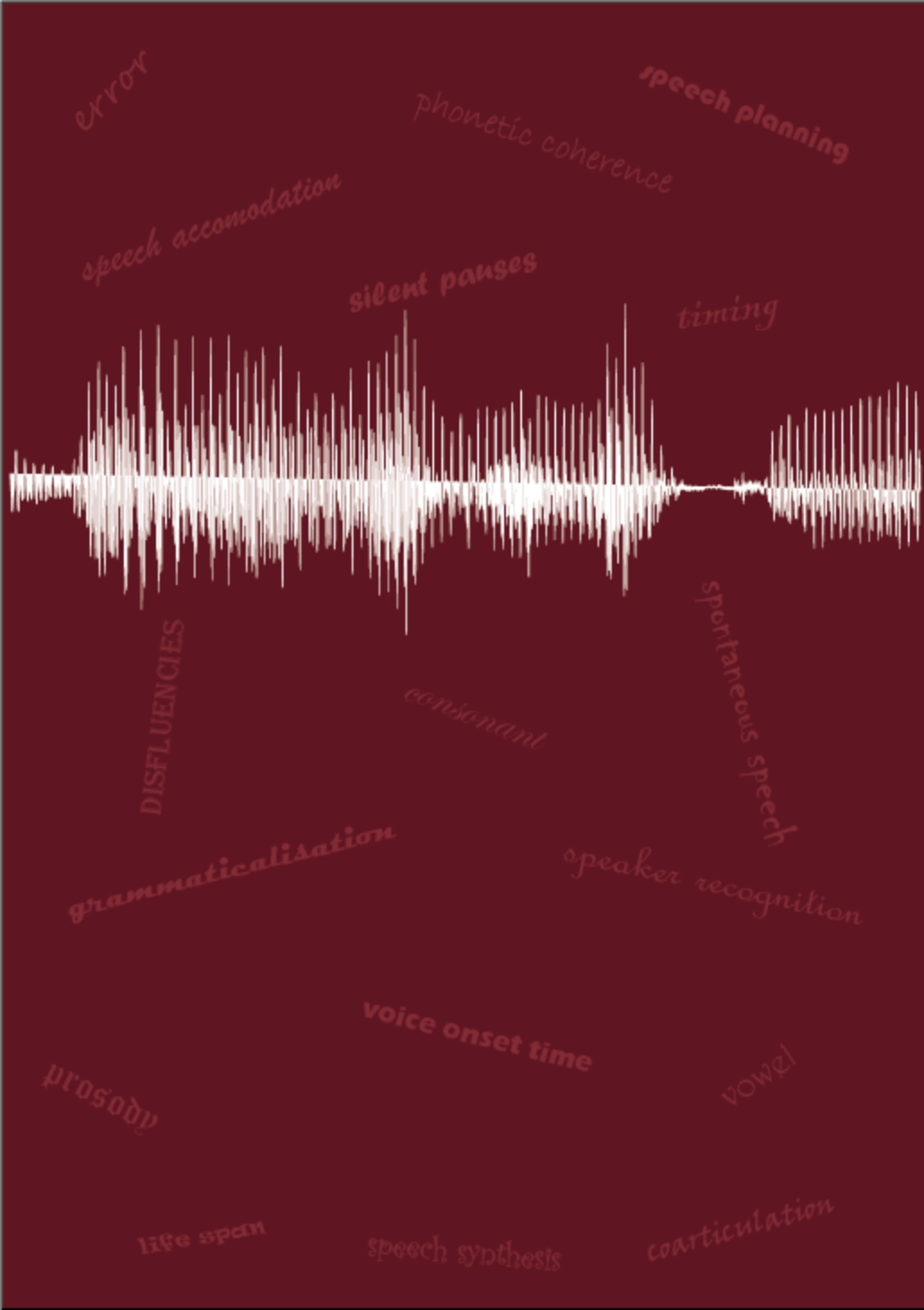
Tigue, C. C., Borak, D. J., O'Connor, J. J. M., Schandl, C., & Feinberg, D. R. (2012). Voice pitch influences voting behavior. *Evolution and Human Behavior, 33,* 210-216.

Tóth, A. (2014). Gyermekek nemének és életkorának meghatározása a beszédük alapján [Estimation of children's age and gender based on their speech]. *Beszédkutatás 2014,* 98-111.

Wayland, R., Herrera, E., & Kaan, E. (2010). Effects of musical experience and training on pitch contour perception. *Journal of Phonetics, 38,* 654-662.

Winkler, R., Brückl, M., & Sendlmeier, W. F. (2003). The aging voice: an acoustic, electroglottographic and perceptive analysis of male and female voices In *Proceedings of the 15th International Congress of Phonetic Sciences* (pp. 2869-2872). 3–9 August 2003. Barcelona.

Winkler, R. (2007). Influences of pitch and speech rate on the perception of age from voice. In *Proceedings of the XVI International Congress of Phonetic Sciences* (pp. 1849-1852). Saarbrücken 6-10 August 2007.


error
phonetic coherence
speech planning
speech accomodation
silent pauses
timing
DISFLUENCIES
consonant
spontaneous speech
grammaticalisation
speaker recognition
voice onset time
prosody
vowel
life span
speech synthesis
coarticulation