Acta Linguistica Hungarica, Vol. 57 (2–3), pp. 329–353 (2010)
DOI: 10.1556/ALing.57.2010.2–3.8

AUTOMATIC RECOGNITION OF SCHWA VARIANTS IN SPONTANEOUS HUNGARIAN SPEECH

ANDRÁS BEKE[a] – GYÖRGY SZASZÁK[b]

[a]Research Institute for Linguistics
Hungarian Academy of Sciences
Benczúr u. 33., H–1068 Budapest, Hungary
beke_andras@freemail.hu

[b]Laboratory of Speech Acoustics
Dept. of Telecommunication and Media Informatics
Budapest University of Technology and Economics
Magyar tudósok körútja 2., H–1117 Budapest, Hungary
szaszak@tmit.bme.hu

Abstract: This paper analyzes the nature of the process involved in optional vowel reduction in Hungarian, and the acoustic structure of schwa variants in spontaneous speech. The study focuses on the acoustic patterns of both the basic realizations of Hungarian vowels and their realizations as neutral vowels (schwas), as well as on the design, implementation, and evaluation of a set of algorithms for the recognition of both types of realizations from the speech waveform. The authors address the question whether schwas form a unified group of vowels or they show some dependence on the originally intended articulation of the vowel they stand for. The acoustic study uses a database consisting of over 4,000 utterances extracted from continuous speech, and recorded from 19 speakers. The authors propose methods for the recognition of neutral vowels depending on the various vowels they replace in spontaneous speech. Mel-Frequency Cepstral Coefficients are calculated and used for the training of Hidden Markov Models. The recognition system was trained on 2,500 utterances and then tested on 1,500 utterances. The results show that a neutral vowel can be detected in 72% of all occurrences. Stressed and unstressed syllables can be distinguished in 92% of all cases. Neutralized vowels do not form a unified group of phoneme realizations. The pronunciation of schwa heavily depends on the original articulation configuration of the intended vowel.

Keywords: vowel neutralization, stressed vs. unstressed syllables, continuous speech, Hungarian, automatic recognition

1. Introduction

The term *vowel reduction* refers to articulatory changes of vowels resulting in a neutralized vowel production that replaces the originally intended

vowel quality. These changes are associated with decreased stress, sonority, duration, loudness, or less articulatory effort. Connected speech, however, requires a dynamic view on vowel production that also takes the influence of phonetic context into consideration. In Bergem (1994) it is shown that the F2-tracks of schwas in various phonetic contexts move almost straight from the onset to the offset, suggesting an articulatory path that requires a minimal amount of effort. To put it differently, schwa is completely assimilated to its phonetic context. Speakers show a tendency to pronounce vowels in a "schwa-like" manner in order to enhance the economy of articulatory gestures in connected speech. In terms of acoustic features this means that the formant frequencies of vowels shift to a position that schwa would have in an identical phonetic context (*idem.*). During the articulation of schwa, the vocal tract configuration is neutral: the lips are unrounded and the tongue is in a central position. The ideal target configuration of the neutral vowel, as characterized by the first three formants, is: F1 = 500 Hz, F2 = 1500Hz, F3 = 2500Hz (Pickett 1999). The hyper-hypo (H&H) theory (Lindblom 1990) focuses on how the speech production mechanism adapts its performance dynamically in answer to the changing perceptual demands. There are two main claims of the H&H theory: (i) speakers hyper-articulate when listeners require maximum acoustic information, and (ii) they reduce articulatory efforts when listeners can supplement the acoustic input with information from other sources (Figure 1).

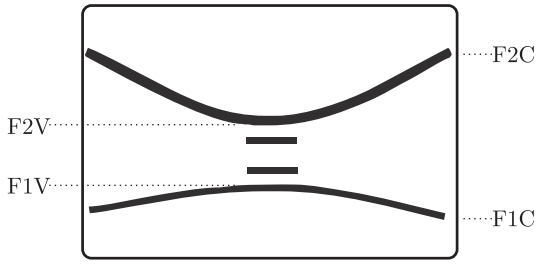

Fig. 1

The schematic figure of schwa's formant structure (from Flemming 2006)

To prevent speakers from over-economizing to a point of unintelligibility, hypo-articulation is governed by a constraint of lexical distinctiveness: speakers hypo-articulate only to the extent that listeners are able to distinguish the target from competing lexical items.

The first two formants of the neutral vowels are assumed to be sufficient for detecting schwa. Various studies found the same correlation between the first two formants and the movement of the tongue into a schwa position. The first formant correlates with the low-high dimension of tongue movement while the second formant correlates with the back-front dimension of tongue movement (Patterson et al. 2003; Slifka 2005; Gósy 2004b; Stevens 1998).

An acoustic tube that is closed at the glottis/posterior end and open at the lips will tend to result in a lower F1 when there is a narrowing of the cross-sectional area in the anterior part of the tube, or a widening of the cross-sectional area in the posterior end of the tube (cf. Stevens 1998). When the tongue body is raised to narrow down the anterior part of the oral tract, the cross-sectional area anterior to the constriction between the tongue dorsum and the palate decreases, and thereby F1 decreases, too. As for the front-back dimension, Stevens showed that forward movement of the tongue body results in higher F2. Several other approaches have suggested that the acoustic realization of schwa depends on diverse factors like phonetic context (Bunnel–Lilley 2008), phonetic positions (Flemming–Johnson 2007), syllable type (stressed vs. unstressed) (Lindblom 1963; Delattre 1969; Gay 1978) and speech style (Masanobu et al. 2006). The acoustic structure of schwa is diverse. Each vowel can be realized as schwa and so the neutralized vowel maintains some spectral features of the original vowels (Koopmans-van Beinum 1994; for Hungarian: Beke 2009) (Figure 2, overleaf). The figure demonstrates the data of the six vowels analyzed ([ɔ] stands for the rounded low back vowel).

Vowels may be neutralized in some of their qualities or functions: in hesitation, in coarticulation (e.g., [r] + schwa), in the "replacing function", etc. (Gósy 2004b; Flemming 2010). In spontaneous speech, schwa may replace various vowels: this will be called the replacing function (7.8% of the vowels are realized as schwa in French, 22.9% in English and 30% in German). In Hungarian, the percentage of schwas replacing various vowels in spontaneous speech is relatively high: 28% (Beke 2009).

Read speech and similar types of speech, e.g., newspaper reading, medical-diagnostic applications or broadcast news, can be recognized (speech-to-text) by a word-accuracy of over 90% using state-of-the-art speech recognition technology. However, recognition accuracy drastically decreases for spontaneous speech (Furui 2007). This decrease is due to the fact that the acoustic and linguistic models used so far are generally built using written language (read speech) that is very different from spontaneous speech (Furui 2005).

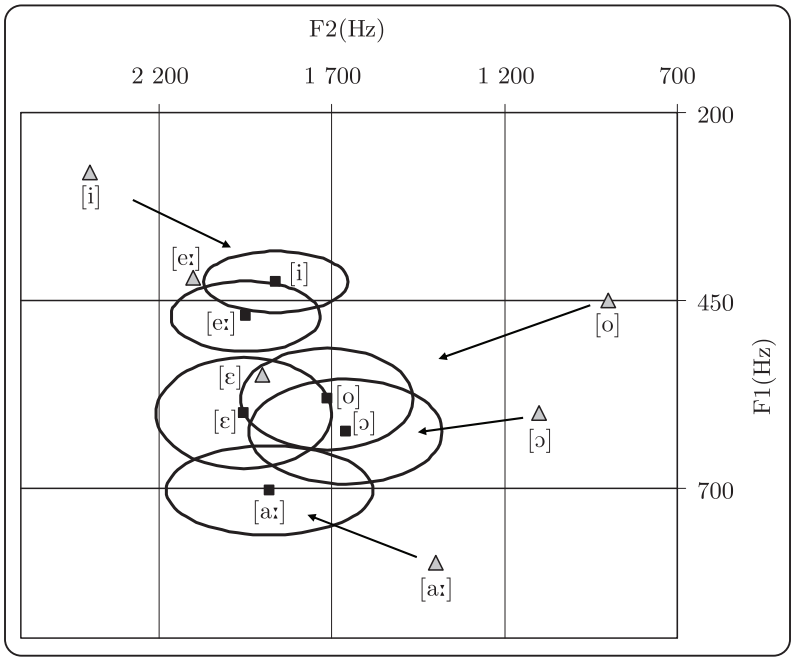

Fig. 2

The medians of the analyzed vowels' space in reading (gray) and in spontaneous speech (black). The ellipses show the ranges of the vowel space in spontaneous speech

Spontaneous speech has significant sources of variation (including phonological variation) and rarely conforms to conventional assumptions and linguistically defined pronunciation rules. All this makes its automatic recognition extremely difficult. Specifically, there may be a lot of different realizations for each expertly defined phonetic unit in the dictionary. The phones may be realized in a clean and complete fashion as in read speech, or they may be realized in a sloppy and incomplete fashion as they often are in spontaneous utterances.

Vowels in unstressed syllables tend to be reduced more than those occurring in stressed syllables. The reduced vowels occurring in unstressed syllables can be used in automatic speech recognition to distinguish stressed vs. unstressed syllables (Xie et al. 2004; Halpern 2006; Lindblom 1963; Delattre 1969; Gay 1978; Koopmans-van Beinum–van Bergem 1989; Engstrand–Nordstrand 1984; Wright–Taylor 1997). The interrelation between vowel reduction and the stressed/unstressed syllable distinction is not perfect since not all unstressed syllables exhibit a reduced vowel (cf. Ladefoged 1993; Janse et al. 2000).

Traditional acoustic-phonetic studies measure formant frequencies of segments to determine their acoustic structures. However, the process of measuring formants is time-consuming, and typically requires careful human judgments to correct formant tracking errors. This problem is exacerbated for segments like schwa that may have very brief duration and less well-defined formant patterns. Because of the time involved, it is practically impossible to analyze large amounts of data. As an alternative to traditional formant-based analyses, in this paper we explore the application of Hidden Markov Models (HMMs) to automatically classify phonetic types and subtypes. The acoustic pre-processing of speech yields Mel-Frequency Cepstral Coefficients (MFCCs) that are used as the input vectors of HMMs both for training and then for classification. Unlike formant frequencies, MFCCs can be calculated quickly and automatically without any human intervention (Bunnel–Lilley 2008).

Our main objective was to carry out the automatic recognition of schwa-realizations based on a large speech corpus of spoken Hungarian (BEA), including a large number of lexical items, and thus phonetic contexts, various speakers, and multiple occurrences. The first question is whether schwa differs in its spectral properties from the basic realizations of the intended vowel phonemes. Our second research question concerns the debate whether schwas phonetically form a relatively unified group of vowels, or the production of schwa is influenced by the (originally intended) vowels. The third question is whether the realization of schwa depends on the actual phonetic context and on stressed vs. unstressed syllables in Hungarian.

2. Subject, database, and method

In this study the spontaneous speech (quasi monologue) of 19 native Hungarian speakers (11 male and 8 female) is used from BEA (***BE****szélt nyelvi* ***A****datbázis* 'spoken language data base' in Hungarian, cf. Gósy 2008). In BEA, the utterances are recorded under silent chamber conditions using a microphone connected to a computer. Goldwave software is used to record the utterances. The sound files are saved in WAVE format at 44000 kHz sampling rate and 16-bit PCM quantization. The phonetic transcriptions of all records were aligned with the speech waveform using Praat software for Speech Analysis. During the analysis, the authors used the following vowels: [ɔ], [aː], [ɛ], [eː], [i], [iː], [o], [oː], [u], [uː]. Segmentations and alignments were carried out manually and controlled both visually

and auditorily. For the identification of schwas, the following strategy was used: (i) the vowel should exhibit centralized formant structures defined in an earlier study by Beke and Gráczi (2009),[1] cf. Figure 3, and (ii) both authors of the present paper should judge their quality as a neutral vowel. In case of disagreement, a third trained person was asked to contribute (this was necessary only in 3.4% of all cases).

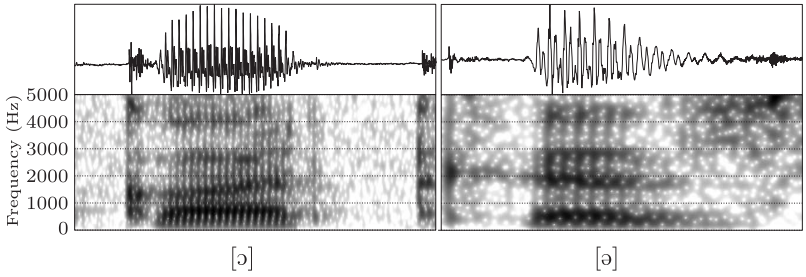

Fig. 3
Spectrograms of the vowels [ɔ] and [ə] (same speaker)

Stressed and unstressed syllables were also annotated. In Hungarian speech, words are always stressed on the first syllable (if at all, cf. Kálmán–Nádasdy 1994). We have double-checked stressed syllables by means of both their F0 and intensity values (Figure 4).

Schwas were identified either as a relatively unified group of neutral vowels, or as various schwa-like vowel qualities replacing Hungarian vowels in spontaneous speech. In the analysis we processed 4,000 vowels in order to devise methods for the recognition of neutral vowels. Mel-Frequency Cepstral Coefficients (MFCC) were calculated and used for the training of Hidden Markov Models (HTK implementation). The recognition system was trained on 2,500 utterances while testing was done on 1,500 further utterances.

2.1. Hidden Markov Models

In automatic speech recognition, Hidden Markov Models (HMM) are commonly used to model the phonemes of a language. In a speech recognition system, a dictionary specifies the pronunciation of words (dictionary

[1] They defined the formant frequencies of the most frequent Hungarian vowels and also schwas that replaced them by means of the K-means algorithm. Schwas were controlled also by auditory inspection.

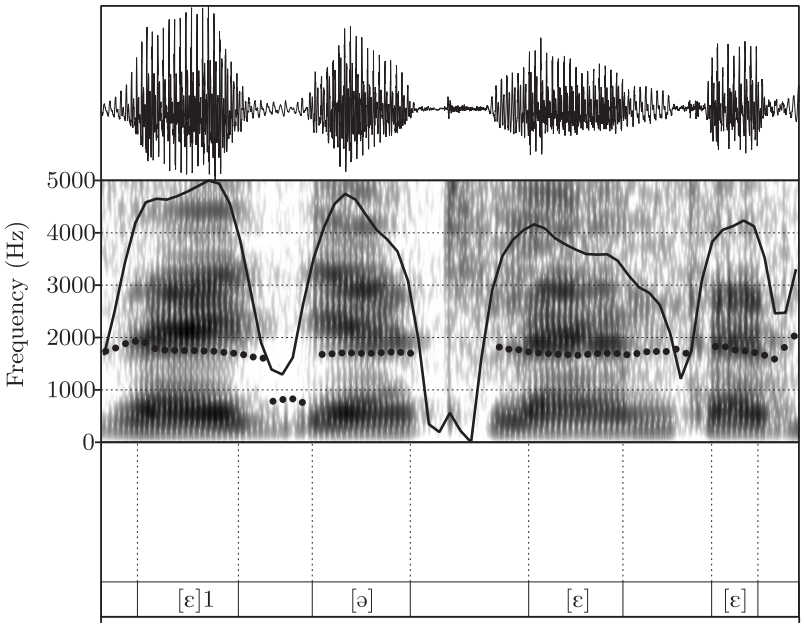

Fig. 4
Stressed [ɛ]1 and unstressed [ɛ], [ə] vowels in Hungarian speech

entries) in the form of phoneme sequences, and a so-called language model specifies which word can follow a given word or word chain. The role of phoneme models is to map speech waveforms to phonemes. To do this, the pure waveform needs to be pre-processed. A frequently used acoustic pre-processing method is the computation of Mel Frequency Cepstral Coefficients (MFCC). The computation of MFC coefficients is as follows: first, a Fast Fourier Transformation (FFT) is applied to the speech waveform. Frequently, a 25-ms part of the speech sample is selected and weighted by a window-function (e.g., Hamming window). Then the window is shifted by the frame rate (usually 10 ms), and another FFT is done. In this way, a speech spectrum is obtained at every 10 ms. The second step of the pre-processing is the decomposition of the spectra corresponding to the critical bands of the human auditory system. This is done by a filterbank (e.g., a Mel filterbank) consisting of 20 separate band-pass filters. Each filter outputs the averaged energy in the given frequency domain covered by the filter. In this way, 20 values in each 10 ms can be obtained. The logarithm of these is taken and a Discrete Cosine Transform (DCT) is applied in order to de-correlate these values

and reduce the dimensions to 12. This means that at this step 12 values—which form a vector or a so-called *frame*—represent each 10 ms of speech. Finally, by adding mean energy and calculating first and second order deltas, one obtains 39-dimensional feature vectors for each 10 ms.

The phoneme HMMs model the distribution of the feature vectors that are assumed to be phoneme-specific. Phoneme HMMs are usually 3-state left-to-right HMMs in order to handle some coarticulation, too. Each state is assigned a probability density function, composed from a weighted mixture of normal distributions (Gaussians) that characterize the "shape" of the feature vectors corresponding to the state (Figure 5).

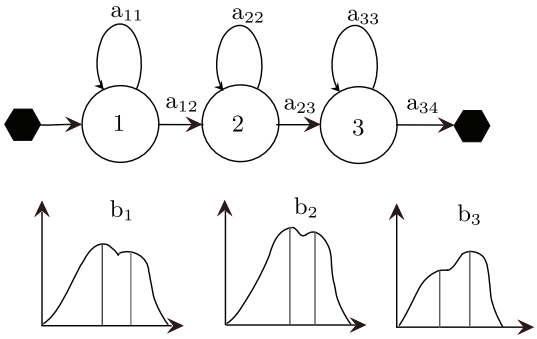

Fig. 5

The structure of a standard 3-state left-to-right phoneme HMM.
Parameters of the model to be estimated are transition probabilities (a_{ij});
weights, means and variances of each Gaussian in output probability functions ($b_j(t)$).

During training, the parameters of these functions are estimated. When used for speech recognition, the feature vectors obtained by the same acoustic pre-processing are compared to the distributions estimated by the mixture. The more they fit, the higher the score of the actual state (sequence) will be when looking for the most probable hypothesis.

Indeed, HMMs in speech recognizers perform a classification task and an alignment task (they classify the phoneme realizations and detect their start and end points). The very same approach can be used to align a phoneme sequence to the input speech. In this case, phoneme classification and phoneme sequence alignment are performed in parallel, this is called *phoneme recognition*. However, this approach can be further simplified by implementing a pure phoneme classification system where phoneme sequence alignment is not needed as each phoneme is presegmented and classified separately. This task is called simply *phoneme*

classification. Both for phoneme recognition and classification, phonemes and/or phoneme classes should be selected for modeling and then, for each class, the HMM should be trained using a statistically representative set of samples. Beyond the trained models, the recognition or classification task also needs a dictionary and a so-called grammar, which is a network or a finite state transducer composed from HMMs. In case of pure phoneme recognition/classification, the dictionary is not necessary and hence, the grammar specifies simply what kind of phoneme or phoneme-class sequences are allowed to be aligned to the input speech (*phoneme recognition*) or what are the classes used for the classification (*phoneme classification*).

A block diagram of the integrated phoneme recognition/classification system used in all experiments is shown in Figure 6. This system was implemented using the HTK toolkit. It is the grammar that specifies whether the system is used for recognition or simply for classification. In the case of phoneme classification, the input speech should be pre-segmented (dotted line in Figure 6) However, as it can be seen, the procedure is quite similar in the two cases.

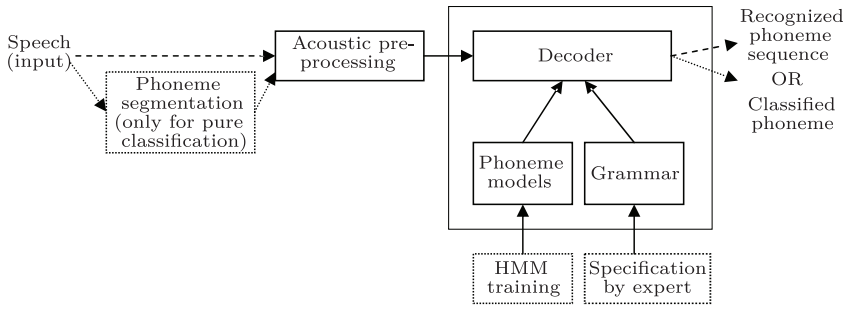

Fig. 6

The integrated phoneme recognition/classification system. (Dashed line: steps needed exclusively for phoneme recognition; dotted line: steps needed exclusively for phoneme classification; normal line: steps necesarry in both tasks.)

3. Results

3.1. Original vowel qualities vs. reduced vowels

Variability in the production of reduced vowels (schwas) seems to be greater than in that of any other vowel (Browman–Goldstein 1992). In

modeling segmental or prosodic phonological processes (Dressler 1984; Madelska–Dressler 1996), schwas may play an important role as the end product of processes such as centralization or in detecting word onsets, thus the separation of full vowels and schwas may be an important goal in speech recognition. A phoneme classification task was designed to analyze the separability of all full vowels merged "V" and all schwas merged "S". This means that "V" was a merged model for all full vowels and "S" was a merged model for all schwa realizations. 3-state left-to-right models (see Fig. 5) were trained using 2, 4, 8, 16, 32 Gaussians in output probability density functions. The grammar used for decoding allowed for both of "V" and "S" with equal weights (probabilities). The best classification result was yielded by the 4 Gaussian models. The results are shown in Table 1.

Table 1

Classification of full vowels and schwas
(merged models, Hungarian, 4 Gaussians)

Vowels	Total	Correct %
Full vowels	706	79.46
Schwas	**157**	**71.97**

Schwas were classified correctly in 71.97% of all schwa realizations. In spontaneous speech, the acoustic realization of schwa is largely variable and the same applies also to full vowels. Based on the investigation of the first and the second formant values, it was confirmed that spontaneous speech is characterized by greater variability of the acoustic properties of vowels than read speech (e.g., Bondarko et al. 2003). In our study the first two formants of the vowels were measured in order to represent the deviations of articulation among vowel realizations. Figure 7 shows a considerable overlap between the full vowels and the reduced vowels (schwas).

These results may contribute to the improvement of the quality of automatic speech recognition systems. Kopecký et al. (2008) proposed to achieve improvement by integrating a "schwa" phoneme into the decoding network for better recognition.

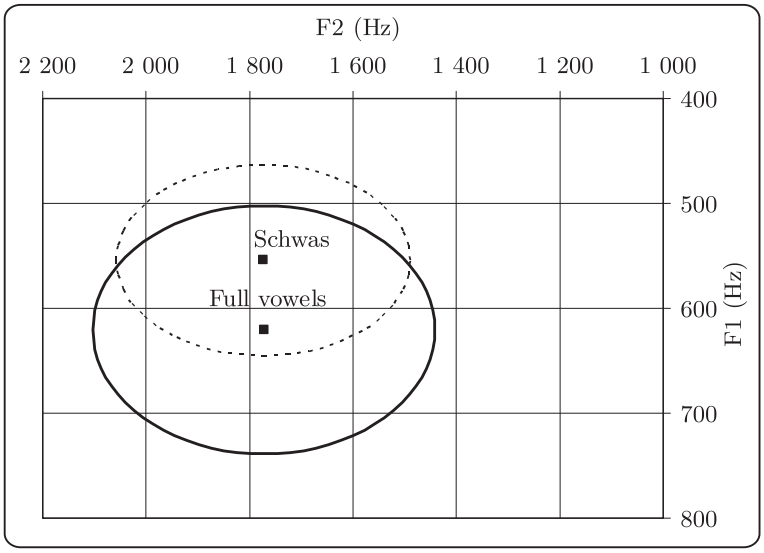

Fig. 7
Plot of full vowel and schwa data for all speakers

3.2. Vowels and merged schwa

Flemming (2010) found that within the articulation gestures of schwas there were two different subtypes: true mid central vowels and contextually variable vowels. Variable schwa assimilates to its context, resulting in substantial contextual variation in vowel quality. Given the existence of two distinct types of schwa vowels, we must be cautious in accepting generalizations about schwa vowels as a group until we really know what kind of vowels are involved. For example, it is clear that the two types of schwa vowel patterns behave quite differently with respect to reduced vowels. In practice, they are transcribed in the same way. Phonological vowel reduction involves the neutralization of vowel contrasts in unstressed syllables, as it is the case, for example, in English. Both mid central and variable schwas arise through vowel reduction but mid central schwa is generally the unstressed counterpart of a low vowel, and arises in a moderate form of vowel reduction that does not affect all vowel qualities, but leaves mid central schwa contrasting with higher vowels. For example, in

Girona Catalan, there are six vowels in stressed syllables [i, e, ɛ, a, o, u] while the vowel inventory is reduced to three vowels [i, ə, u] in unstressed syllables where schwa is a mid central vowel (Herrick 2003). The vowels [e, ɛ, a] are reduced to [ə] in unstressed syllables, while [o, u] neutralize to [u]. Variable schwa results from a more extreme form of vowel reduction that applies to all vowel qualities, potentially neutralizing all vowel qualities, like in English.

The preceding experiment (the classification of schwas vs. full vowels) has shown that full vowels (in general) and schwas can be well separated. In the next experiment, a 3-state HMM is constructed for the individual vowels [ɔ], [ɛ], [o] and [ə] (all schwas merged) labels in the transcription set. The best recognition results for these four different vowels were obtained by 8 Gaussian models (see Table 2) in this phoneme recognition task.

Table 2

Recognition results for vowels [ɔ], [ɛ] and [o], and for vowel-independent (merged) schwa [ə]

Vowels	Total	Correct %	Correct % without deletion %	Deletion[2]
[ə]	**140**	**65.00**	**65.30**	8
[ɔ]	167	70.65	70.95	14
[ɛ]	225	75.11	75.36	8
[o]	115	73.04	73.51	8

In Table 3, the confusion matrix[3] for the experiment is shown. As can be seen, the vowel [o] is somewhat more frequently involved in confusions than [ɔ], and especially than [ɛ].

[2] There are three types of errors in recognition tasks: deletion, insertion, and substitution. A deletion error occurs if the recognizer misses a phoneme. (It does not identify it as a separate phoneme when aligning the phoneme sequence to the input speech. In classification, however, only substitution errors may occur.) If one discards deletion errors, a ratio is obtained which can be interpreted as classification performance; however, in this case the missed phonemes are excluded from evaluation, distorting the results compared to "classical" classification. In other words, "correct without deletion" rate is the classification rate of the identified phonemes.

[3] Confusion matrices are used to analyze substitution errors (both in classification and recognition tasks).

Table 3
Confusion matrix for vowels [ɔ], [ɛ] and [o], and for vowel-independent (merged) schwa [ə]

Vowels	[ɔ]	[ə]	[ɛ]	[o]
[ɔ]	77.12	5.88	6.54	**10.46**
[ə]	11.36	**68.94**	**6.06**	**13.64**
[ɛ]	7.37	5.99	77.88	**8.76**
[o]	11.21	6.54	3.74	78.50

The reason for this, in our opinion, is that the articulation configuration of the vowel [o] is more similar to that of schwa; moreover, the duration of the vowel [o] is less than that of [ɛ] (Gósy 2004a;b). The duration of vowels and schwas were compared, and a significant difference was found between them. The duration of schwa is significantly shorter than the duration of the original vowels. The mean duration of schwa was 53 ms while the mean duration of the vowels was 84 ms (ANOVA: $F(1, 2917) = 252.757$, $p = 0.000^{**}$) (see Figure 8). This difference seems to be a language-independent fact (Bondarko et al. 2003; Gósy 2004b; Flemming 2010; Swerts et al. 2007).

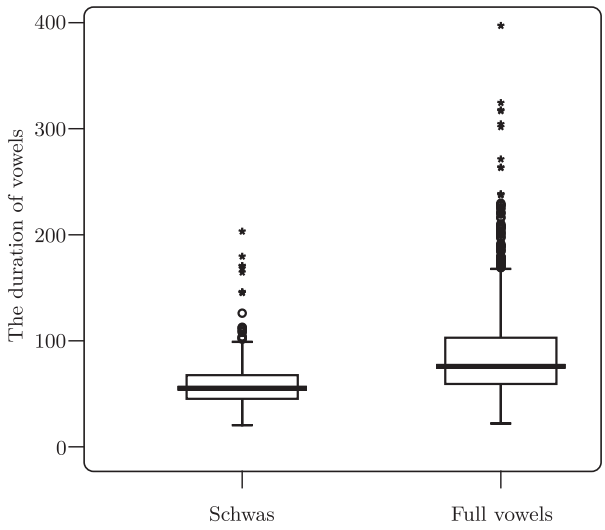

Fig. 8
The durations of the original vowels and the neutralized vowels

Figure 9 shows the temporal differences among the three vowels. The vowel [o] (77 ms) is significantly shorter than [ɛ] (83 ms) or than [ɔ] (90 ms) (ANOVA: $F(2, 2313) = 19.86$, $p = 0.000^{**}$; the between subject (post hoc test) $p > 0.000^{**}$).

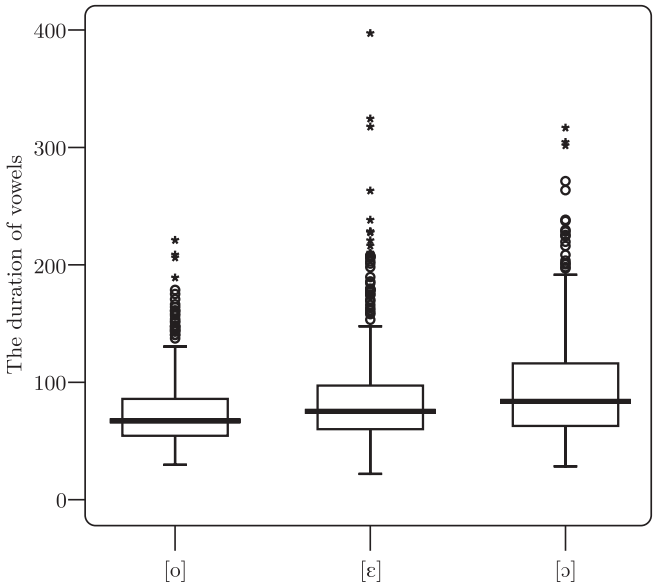

Fig. 9
The durations of the vowels [o], [ɛ] and [ɔ]

The decrease in vowel duration is caused by the phenomenon of undershoot (cf. Lindblom 1990). Target undershoot is responsible for numerous variations in vowel realizations. Bondarko et al. (2003) evaluated the results of a systematically performed comparative study of Russian read vs. spontaneous speech. As expected, the statistical results showed greater variability in vowel durations in spontaneous speech than in read speech: both variance and standard deviation values were higher in spontaneous speech.

The more vowel realizations in spontaneous speech vary, the broader the interval of the formant values. The first and second formant values of the vowels analyzed also confirm that spontaneous speech is characterized by a greater variability of the acoustic properties of vowels (Figure 10).

The vowel transitions can be identified in all vowels both in reading and in spontaneous speech while the steady-state parts were often absent from spontaneous speech as the vowels in this case were more reduced

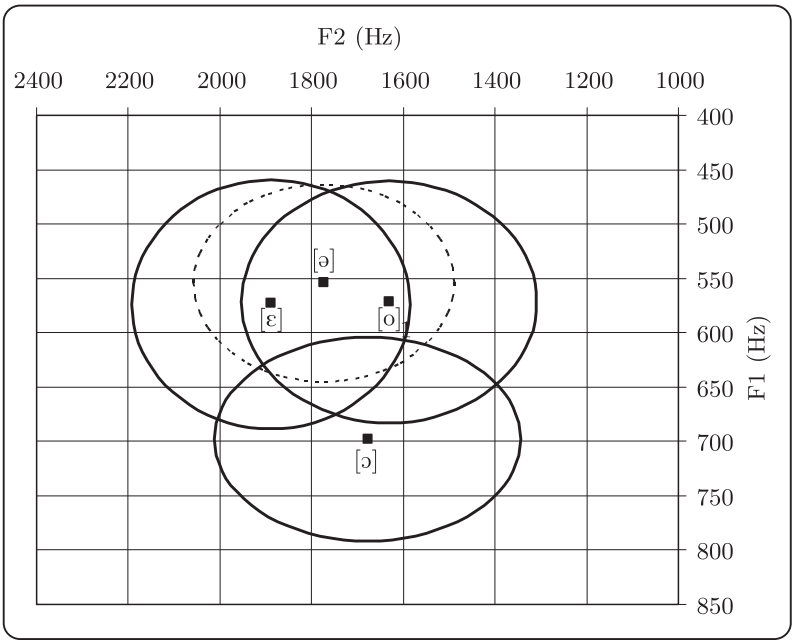

Fig. 10
Plot of vowel and schwa data for all speakers

(Bondarko et al. 2003). Since the vowel [o] is shorter then [ɔ] and [ɛ], its steady-state part is often missing. Our results correspond to those of Padgett and Tabain (2005). In a study on the read speech of nine speakers, these authors examined phonological vowel reduction in Russian. They found that the vowels [a] and [o] could be poorly distinguished from each other both in stressed and unstressed syllables. The vowel [o] was correctly identified better than chance but [a] was more often identified as [o] than as [a].

3.3. Vowels and vowel-dependent schwas

Phonetic vowel reduction is a consequence of some modified articulatory gestures like decreased stress, less sonority, shorter duration, less loudness and less articulatory effort, respectively. However, there is no general picture about the nature of vowel reduction (cf. Koopmans-van Beinum 1994).

3-state HMMs were constructed for each vowel [ɔ], [ɛ] and [o], and for each schwa **[A]**, **[E]**, **[O]**. These schwa models differ from their original vowels, [ɔ], [ɛ] and [o], respectively, so they are vowel-dependent schwa models. The schwas form a well separable group of vowels as opposed to the original basic realizations of the phonemes in question. However, the group of schwa vowels is not as integrated as the other groups of various vowels. The reason for that—as was mentioned before—is that schwas preserve some original gestures (as seen in their spectrum) of the vowels they are to replace.

A garbage model (G) is also trained to cover all other vowels not modeled explicitly. Each model was trained using 2, 4, 8, 16, 32 Gaussians in the output probability density functions. The grammar used for decoding allowed for all the 7 types of phonemes to occur with equal weights (probabilities). The best result was yielded by the 4 Gaussian models (see Table 4).

Table 4
Recognition results for the vowels [ɔ], [ɛ] and [o], and for vowel-dependent schwas **[A]**, **[E]**, **[O]**

Class	Total	Correct %	Correct without deletion %	Deletion
[ɔ]	169	65.08	69.62	11
[A]	47	**68.08**	**72.72**	3
[ɛ]	227	69.60	71.49	6
[E]	65	**63.07**	**68.33**	5
[o]	116	61.20	61.73	1
[O]	29	**62.06**	**64.28**	1

Table 4 shows that it is schwas (**[A]**) replacing [ɔ] vowels that provided us with the best recognition rates (68%, 72.72%). On the contrary, schwas (**[O]**) replacing [o] vowels gave the worst performance (62.06% and 64.28%). The reason for the poor results of the latter vowels lies in their extremely short duration that hamper the operations of recognition. The recognition results are better using this model than they were using the merged model of schwa. What is more important is that this result confirms our hypothesis of the vowel dependency of schwas.

3.4. Front and back vowels vs. front and back schwas

It has been found that the acoustic realizations of schwa appear in three functions in numerous languages (cf. Gósy 2004b): hesitation, coarticulation and the replacing function. Our hypothesis is that the acoustic properties of the neutral vowels in the replacing function depend on the articulation configuration of the original vowels. In other words, we assumed that schwa realizations preserve specific patterns of the original vowels. In this study we examined the acoustic properties of schwa realizations in the replacing function. Several studies have already suggested that the second formant values and the vowel durations were the basic acoustic parameters of neutralization. This claim involves that two subtypes of schwa can be separated in the acoustic dimensions, corresponding to front and back schwas in articulation. Lilley (2008) trained a set of 55 monophones using Hidden Markov Models which were automatically extracted and tested on a phonetically diverse corpus of 1,837 utterances from a single adult female talker. He supposed that there would be a distinction between "front" and "back" schwas. He found that the clusters consisted almost exclusively of either front schwa models or back schwa models, suggesting that the difference between "front" and "back" schwas is really manifested in this speaker's pronunciation. However, the two schwa types were found mixed in many clusters, a fact that indicates that though the onsets of these vowels are distinct, their offsets are quite similar.

Figure 11 (overleaf) shows that back vowels and the corresponding "back" schwas are more different from one another than front vowels and corresponding "front" schwas are.

Again, 3-state HMMs were constructed for both back vowels and front vowels, and for both back schwas and front schwas. 4 Gaussian models gave the best recognition results (see Table 5).

Table 5

Recognition results for front and back vowels, and for front and back schwas

Vowels	Total	Correct %	Correct without deletion %	Deletion
Back vowels	318	56.91	71.54	65
Front vowels	375	53.86	68.70	81
Back schwa	76	**63.15**	**73.84**	11
Front schwa	66	**40.90**	**54**	16

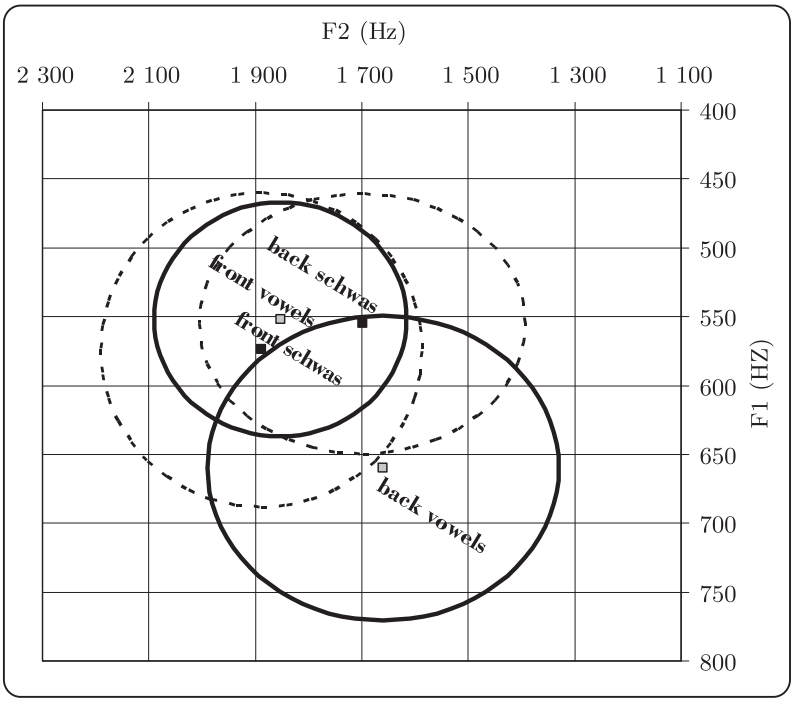

Fig. 11

Plots of back and front vowels and back and front schwa data, for all speakers

These results support previous descriptions of schwas that suggest there exist distinct forms that differ in both back and front variants. Table 5 shows that back schwas are identified at the highest rate. This can be explained by the fact that the acoustic realization of front vowels and front schwas is less homogeneous, thus the models were less capable of identifying them. This result corresponds to that of Bunnel and Lilley (2008). They found that the rate of correct recognition was better in the case of back schwas than in that of front schwas.

In the second part of the experiment, 3-state HMMs were constructed for both back schwas and front schwas in order to investigate how efficiently they can be separated automatically. 4 Gaussian models gave the best classification results (see Table 6).

The results show again that back schwas form a more homogeneous class than front schwas. For this reason, the classification of back schwas is slightly better. Our results support the claim (Flemming–Johnson 2007) that "front schwa" is more common than it has generally been assumed.

Table 6
Classification results for front and back schwas

	Total	Correct %
Back schwa	89	79.77
Front schwa	69	66.66

This indicates that schwa in its replacing function has two well-isolable subtypes: back schwas and front schwas depending on the original vowel's articulation configuration. It has been found that the centralization was highly significant in the cases of [i], [ɛ], [o] and [u] whilst it shows weak significance in the case of [ɔ]. In addition, the δ values ($\delta =$ centralization index: it is defined as the difference between the laboratory speech sound distance to schwas and the spontaneous speech sound distance to schwas) indicate a more marked centralization in front than in back vowels (Harmegnies–Poch-Olivé 1992).

3.5. Stressed vs. unstressed syllable detection on the basis of vowel quality

Hungarian, a language of the Finno-Ugric family, is an agglutinating language characterized by a relatively free word order (Siptár–Törkenczy 2000). It is a syllable-timed language where word stress always falls on the first syllable of words. Due to the different structural characteristics of the two languages, standard methods developed for English automatic speech recognition cannot be directly applied (Szaszák–Vicsi 2007).

As it is well known, vowel quality is determined by the configuration of the tongue, jaw, and lips (Ladefoged 1993; Bernthal–Bankson 1998; Ladefoged–Maddieson 1996; Pennington 1996, etc.). The flexibility of vowel articulation leads to the variability of realizations of the same phoneme. The actual articulation configuration of a vowel may have an indication about bearing stress or not. Reduced vowels tend to have a more central articulation than full vowels: i.e.,, the tongue is closer to its "rest" position (cf. Cruttenden 1997; Ladefoged 1993). If the syllable is unstressed, other vowels may also be pronounced in reduced forms, making them more schwa-like (Swerts et al. 2007).

3-state HMMs were constructed for vowels in stressed syllables "XA" and unstressed syllables "XT". Two models were developed in order to

distinguish the stressed and unstressed vowels on the one hand, and the stressed, unstressed and schwa vowels, on the other hand. Each model was trained using 2, 4, 8, 16, 32 Gaussians in the output probability density functions. The grammar used for decoding allowed for the two classes to occur with equal weights (probabilities). The best result was yielded by the 8 Gaussian model (Table 7).

Table 7

Stressed vs. unstressed syllable recognition based on vowel quality

	Total	Correct %	Correct without deletion %	Deletion
XA	309	**82.80**	95.93	39
XT	855	**70.72**	91.87	195

The ratio of correct recognition was 73.76% while correct recognition without automatic system deletions was 92.98% (this result can be interpreted as a pure classification result where the missed phonemes are excluded from the testing). The recognition of the stressed syllables was better (82.80%) than that of the unstressed syllables (70.72%). Correct classification without automatic system deletions is 95.93% of all cases (XA) and 91.87% of all cases (XT), respectively. This result is slightly better than Xie et al.'s (2004) (82.50%) for English in a similar study. In Xie et al.'s study, a combination of duration and amplitude features provided the best performance (84.72%) and vowel quality features also provided good results (82.50%). The authors noted that both prosodic features and vowel quality features were equally effective at detecting stress but their combination did not increase the classification performance. In our experiment for the stressed/unstressed syllable distinction, only the vowel quality features were used.

The recognition ability for the distinction of unstressed schwas "XS", unstressed vowels "XT" and stressed vowels "XA" was also tested. Each model was trained using 2, 4, 8, 16, 32 Gaussians in the output probability density functions. The grammar used for decoding allowed for all the three classes to occur with equal weights (probabilities). The best result was yielded by the 8 Gaussian models (Table 8).

The ratio of correct recognition is 72.61% while correct classification without automatic system deletion is, as expected, higher, 84.22% in average for all cases. Although the ratio of the correct recognition is

Table 8
Stressed vs. unstressed syllable and schwa recognition based on vowel quality

	Total	Correct %	Correct without deletion %	Deletion
XA	299	80.70	93.11	38
XT	646	68.80	79.52	86
XS	218	73.20	86.44	32

slightly lower than it was obtained in the previous recognition task, the ratio of the recognition of the schwa vowels is the highest compared to the previous data (73.20, 86.44%). (Deletion error rate was the lowest in this case.) The results support the assumption that the schwa realizations depend primarily on the stress position of the vowel in the word.

4. Conclusions

The aim of this paper was to develop an automatic recognition process using HMMs for detecting schwa vowels in spontaneous speech in the function of replacing Hungarian vowels. Spontaneous speech is highly variable and rarely conforms to conventional assumptions and linguistically defined pronunciation rules, a fact that makes its automatic recognition extremely difficult.

We have shown that (i) schwa and its variations are automatically detectable in spontaneous Hungarian; and (ii) schwa variations depend primarily on the articulation configuration of the original vowel. In the first experiment, we trained five HMM models by which we could represent the schwa and its possible subtypes. The aim here was to identify the vowels as either the basic realization of the intended phoneme or a schwa. This model can correctly classify 78% of all (merged) schwas. The best result was gained by the 8 Gaussian models.

The second aim of this study was to gain a better understanding of the nature of schwa. The acoustic configurations of fully pronounced vowels and schwas were compared. The results highlighted the strong similarity of the articulation configuration of [o] and the articulation configuration of schwa. The reason for this is that the articulation configuration and the duration of the vowel [o] show greater variability in spontaneous speech than those of any other vowel.

The third aim of this research concerned the distinction between "front" and "back" schwa subtypes. Is there a clear-cut distinction between these alleged schwa subtypes, or do schwas vary more significantly along some other dimension? Our results support the claim that "front schwa" is more common than it was generally assumed. This indicates that schwa in its replacing function has two subtypes that can be clearly distinguished: back schwa and front schwa, depending on the original vowel's articulation configuration. The neutral vowel is not a homogeneous (unified) vowel category because it inherits, in a way, the articulation configurations of the originally intended vowels. The ratios of correct recognition were better when schwas were separated in terms of the articulation configuration of the originally intended vowels.

In the last part of the study, the role of vowel quality was analyzed in an automatic stress detection process. The ratio of correct recognition is 73.76% of all cases for the distinction of stressed vs. unstressed syllables while correct recognition without automatic system deletion is 92.98% of all cases. The recognition of the stressed syllables gave the best results: 82.80% of all cases. The ratio of correct classification (without automatic system deletions) was 95.93% of all cases. The vowel quality parameters used in our study for the distinction between stressed and unstressed syllables gave acceptable results which are slightly better than those of similar systems in this area (cf. Jenkin–Scordilis 1996; Kuijk–Boves 1999).

References

Beke, András 2009. A veláris magánhangzók stabilitása a spontán beszédben [The invariance of velar vowels in spontaneous Hungarian speech]. In: Tamás Gecső – Csilla Sárdi (eds): A kommunikáció nyelvészeti aspektusai [Linguistic aspects of communication], 27–31. Kodolányi János Főiskola & Tinta Könyvkiadó, Székesfehérvár & Budapest.

Beke, András – Tekla Etelka Gráczi 2009. A magánhangzók semleges realizációja a spontán beszédben [The neutral realization of vowels in spontaneous hungarian speech]. Paper presented at the Eleventh Summer School of Psycholinguistics, Balatonalmádi, 2009.

Bergem, Dick R. van 1994. A model of coarticulatory effects on the schwa. In: Speech Communication 14 : 143–62.

Bernthal, John E. – Nicholas W. Bankson 1998. Articulation and phonological disorders (4th ed.). Prentice Hall, Englewood Cliffs NJ.

Bondarko, Liya V. – Nina B. Volskaya – Svetlana O. Tananaiko – Ludmila A. Vasilieva 2003. Phonetic properties of Russian spontaneous speech. In: Maria-Josep Solé – Daniel Recasens – Joachim Romero (eds): Proceedings of the 15th International

Congress of Phonetic Sciences (Barcelona, 3–9 August 2003), 2973–6. Causal Productions, Barcelona.

Browman, Catherine P. – Louis Goldstein 1992. Articulatory phonology: An overview. In: Phonetica 49 : 155–80.

Bunnel, H. Timothy – Jason Lilley 2008. Schwa variants in American English. In: Proceedings of the InterSpeech 2008 Conference, Brisbane, Australia, 1159–62. International Speech Communication Association, Brisbane.

Cruttenden, Alan 1997. Intonation (2nd ed.). Cambridge University Press, Cambridge.

Delattre, Pierre 1969. An acoustic and articulatory study of vowel reduction in four languages. In: International Review of Applied Linguistics 7 : 295–325.

Dressler, Wolfgang U. 1984. Explaining Natural Phonology. In: Phonology Yearbook 1 : 29–50.

Engstrand, Olle – Lennart Nordstrand 1984. Acoustic features correlating with tenseness, laxness and stress in Swedish: Preliminary observations. In: Claes-Christian Elert – Iréne Johansson – Eva Strangert (eds): Nordic Prosody III. Acta Universitatis Umensis (Umeå Studies in the Humanities 59), 51–66. University of Umeå, Umeå. ([Also in: Reports from Uppsala University, Department of Lingustics (RUUL) 11: 8–22.).

Flemming, Edward 2006. Contrast and schwa vowels in English. Manuscript, MIT. http://web.mit.edu/flemming/www/paper/schwa.ppt.

Flemming, Edward 2010. The phonetics of schwa vowels. In: Donka Minkova (ed.): Phonological weakness in English: From Old Present-Day English, 78–95. Palgrave Macmillan, Basingstoke.

Flemming, Edward – Stephanie Johnson 2007. Rosa's roses: Reduced vowels in American English. In: Journal of the International Phonetic Association 37 : 83–96.

Furui, Sadaoki 2005. Recent progress in corpus-based spontaneous speech recognition. In: IEICE-Transactions on Information and Systems E88-D : 366–75.

Furui, Sadaoki 2007. Recent advances in automatic speech summarization. In: David Evans – Sadaoki Furui – Chantal Soul (eds): Proceedings of the IEEE/ACL. Workshop on Spoken Language Technology, IEEE, Los Alamitos, 2006, 115–22. CID, Pittsburg.

Gay, Thomas 1978. Effect of speaking rate on vowel formant movements. In: Journal of the Acoustical Society of America 63 : 223–30.

Gósy, Mária 2004a. Fonetika, a beszéd tudománya [Phonetics, the science of speech]. Osiris Kiadó, Budapest.

Gósy, Mária 2004b. The manifold function of schwa. In: Grazer Linguistische Studien 62 : 15–26.

Gósy, Mária 2008. Magyar spontánbeszéd-adatbázis – BEA [Hungarian spontaneous speech corpus – BEA]. In: Mária Gósy (ed.): Beszédkutatás 2008 [Speech research 2008], 194–207. MTA Nyelvtudományi Intézet, Kempelen Farkas Beszédkutató Laboratórium, Budapest.

Halpern, Jack 2006. The contribution of lexical resources to natural language processing of CJK languages. In: Quiang Huo – Bin Ma – Chng Eng Siong – Haizhou Li (eds): International Symposium on Chinese Spoken Language, ISCSLP, Singapore, 2006, 768–80. Springer, Berlin.

Harmegnies, Bernard – Dolors Poch-Olivé 1992. A study of style-induced vowel variability: Laboratory versus spontaneous speech in Spanish. In: Speech Communication 11 : 429–437.

Herrick, Dylan 2003. An acoustic analysis of phonological vowel reduction in six varieties of Catalan. Doctoral dissertation, University of California, Santa Cruz.

Janse, Esther – Anke Sennema – Anneke Slis 2000. Fast speech timing in Dutch: The durational correlates of lexical stress and pitch accent. In: Proceedings of the VIth International Conference on Spoken Language Processing, Beijing, October 2000, vol. III, 251–4. International Speech Communication Association, Beijing.

Jenkin, Karen L. – Michael S. Scordilis 1996. Development and comparison of three syllable stress classifiers. In: Quiang Huo – Bin Ma – Chng Eng Siong – Haizhou Li (eds): International Symposium on Chinese Spoken Language, ICSLP, 1996, 733–6. Springer, Singapore.

Kálmán, László – Ádám Nádasdy 1994. A hangsúly [Stress]. In: Ferenc Kiefer (ed.): Strukturális magyar nyelvtan 2: Fonológia [A structural grammar of Hungarian 2: Phonology], 393–467. Akadémiai Kiadó, Budapest.

Koopmans-van Beinum, Florina J. 1994. What's in a schwa? Durational and spectral analysis of natural continuous speech and diphones in Dutch. In: Phonetica 51 : 68–80.

Koopmans-van Beinum, Florina J. – Dick R. van Bergem 1989. The role of 'given' and 'new' in the production and perception of vowel contrasts in read text and in spontaneous speech. In: EUROSPEECH 1989 : 2113–6.

Kopecký, Jiří – Ondrej Glembek – Martin Karafiat 2008. Advances in acoustic modeling for the recognition of Czech. Paper presented at the International Conference on Text, Speech and Dialogue, TSD 2008.

Kuijk, David van – Loe Boves 1999. Acoustic characteristics of lexical stress in continuous telephone speech. In: Speech Communication 27 : 95–111.

Ladefoged, Peter 1993. A course in phonetics. Harcourt, Brace and Jovanovich, Fort Worth, Texas.

Ladefoged, Peter – Ian Maddieson 1996. The sounds of the world's languages. Blackwell, Cambridge MA & Oxford.

Lilley, Jason 2008. Data-driven investigation of subphonemic variation: "Front" schwa vs. "back" schwa. Paper presented at the Cognitive Science Graduate Student Conference 2008, Delaware, April 18th, 2008.

Lindblom, Björn 1963. Spectrographic study of vowel reduction. In: Journal of the Acoustical Society of America 35 : 1773–81.

Lindblom, Björn 1990. Explaining phonetic variation: A sketch of the h&h theory. In: William J. Hardcastle – Alain Marchal (eds): Speech production and speech modeling, 403–40. Kluwer, Dordrecht.

Madelska, Liliana – Wolfgang U. Dressler 1996. Postlexical stress processes and their segmental consequences illustrated in Polish and Czech. In: Bernhard Hurch – Richard A. Rhodes (eds): Natural Phonology: The state of the art, 189–200. Mouton de Gruyter, Berlin & New York.

Masanobu, Nakamura – Furui Sadaoki – Iwano Koji 2006. Acoustic and linguistic characterization of spontaneous speech. In: Renato de Mori (ed.): Speech Recognition

and Intrinsic Variation Workshop, SRIV, Toulouse, France, May 20, 2006, 3–8. International Speech Communication Association, Toulouse.

Padgett, Jaye – Marija Tabain 2005. Adaptive Dispersion Theory and phonological vowel reduction in Russian. In: Phonetica 62 : 14–54.

Patterson, David – Paul C. LoCasto – Cynthia M. Connine 2003. Corpora analyses of frequency of schwa deletion in conversational American English. In: Phonetica 60 : 45–69.

Pennington, Martha C. 1996. Phonology in English language teaching: An international approach. Longman, London.

Pickett, James M. 1999. The acoustics of speech communication: Fundamentals, speech perception theories, and technology. Ally and Bacon, Boston & New York.

Siptár, Péter – Miklós Törkenczy 2000. The phonology of Hungarian. Oxford University Press, Oxford.

Slifka, Janet 2005. Acoustic cues to vowel–schwa sequences for high front vowels. In: Journal of the Acoustical Society of America 118 : 2037.

Stevens, Kenneth N. 1998. Acoustic phonetics. MIT Press, Cambridge MA.

Swerts, Marc – Hanne Kloots – Steven Gillis – Georges De Schutter 2007. Vowel reduction in spontaneous spoken Dutch. In: Sadaoki Furui (ed.): Proceedings of the ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition, Tokyo, Japan, 2007, 31–4. International Speech Communication Association, Tokyo.

Szaszák, György – Klára Vicsi 2007. Speech recognition supported by prosodic information for fixed stress languages. In: Václav Matoušek – Pavel Mautner (eds): Text, speech and dialogue. Proceedings of the 10th TSD conference Pilsen, 262–9. Springer, Heidelberg.

Wright, Helen – Paul A. Taylor 1997. Modelling intonational structure using hidden Markov models. Paper presented at the ESCA workshop on Intonation: Theory, Models, and Applications, Athens, Greece, 1997.

Xie, Huayang – Peter Andreae – Mengjie Zhang – Paul Warren 2004. Detecting stress in spoken English using decision trees and support vector machines. In: James M. Hogan – Paul Montague – Martin K. Purvis – Chris Steketee (eds): ACSW Frontiers '04. Proceedings of the Second Workshop on Australasian Information Security, Data Mining and Web Intelligence, and Software Internalisation, vol. 3, 145–50. Australian Computer Society, Dunedin.