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%

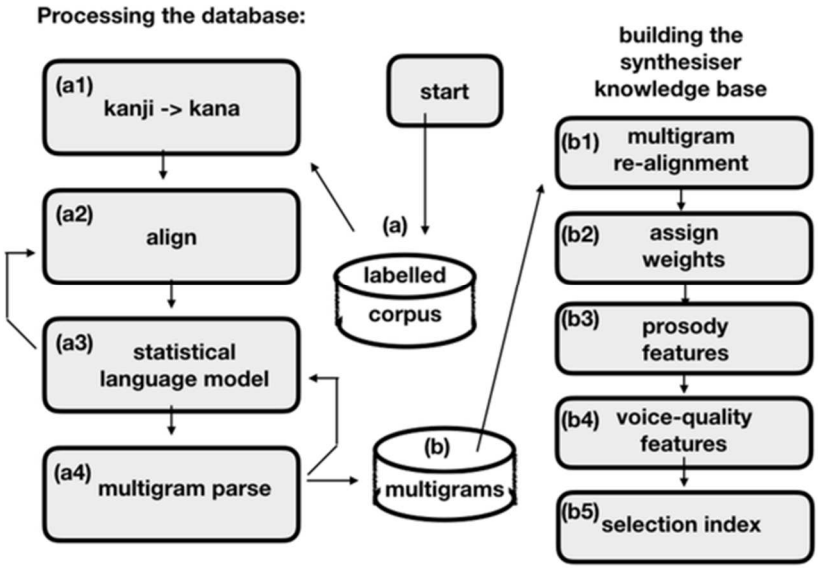

Figure 2.
Flow of processing of the GBO corpus for use in CHATR synthesis

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.

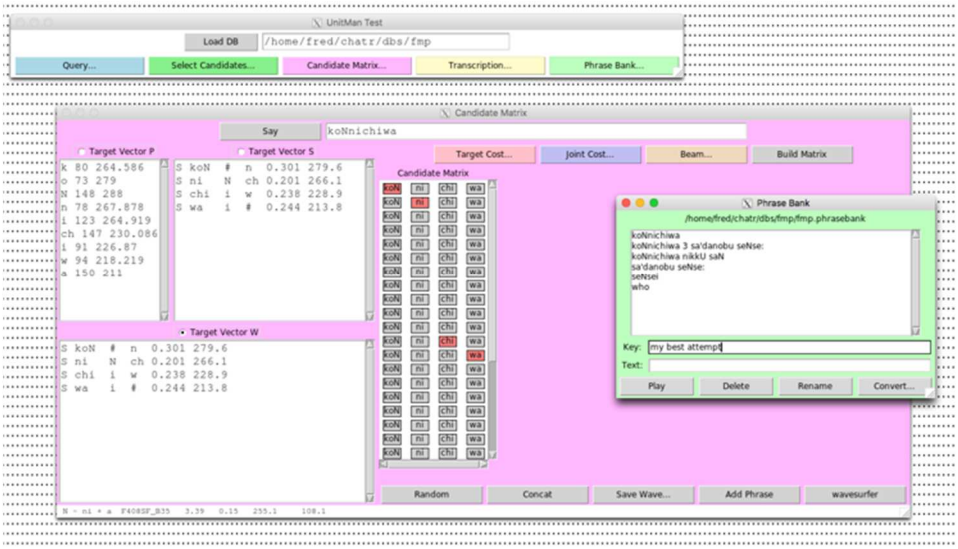

Figure 3.
CHATR's UnitMan and PhraseBank modules

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?

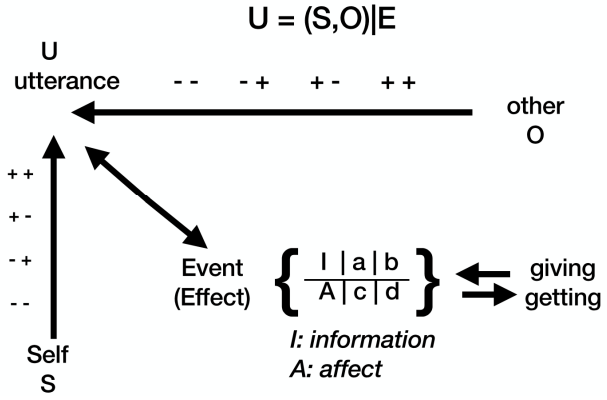

Figure 4.
Criteria for higher-level feature-based unit selection incorporating pragmatic constraints (from: Campbell (2004) User Interface for an Expressive Speech Synthesiser, IEICE Tech Rept.)

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.