INCREASING THE NATURALNESS OF SYNTHESIZED SPEECH

Tamás Gábor Csapó
BME TMIT
e-mail: csapot@tmit.bme.hu

Abstract

In my ongoing PhD work, I am doing research in the field of text-to-speech (TTS) synthesis, particularly statistical parametric speech synthesis. For increasing the naturalness of synthesized speech, three main topics are dealt with: 1) introducing pitch variability, 2) novel excitation modeling and 3) investigating the role of subglottal resonances. 1) By investigating the fundamental frequency of speech, variability of pitch was modeled and the prosody component of a speech synthesizer was improved. Contrary to the traditional deterministic prosody models, we proposed a method to assign several pitch variants not differing in meaning for synthesized sentences, thus making the pitch component of speech synthesis more variable over longer passages. 2) In a separate experiment, the "buzzy" quality of parametric speech synthesis was reduced. Source-filter decomposition was used to obtain the speech residual signal and a novel codebook-based excitation model was proposed for use in the statistical parametric text-to-speech synthesis framework. This method uses phoneme-dependent residual frames which is an improved modeling technique compared to similar methods. 3) The role of subglottal resonances (SGRs) on speech production was investigated. We have shown that the SGRs have important phonological effects in the Hungarian language as well. Our future goal is to model the influence of the subglottal tract in speech synthesis, which is not explicitly modeled in the traditional source-filter model. Our results contribute to make synthesized speech more natural. Variable pitch has been shown to improve the naturalness of synthesized speech in a specific scenario. The novel excitation model can produce similar quality speech to other vocoders, moreover it will be possible to model different voice qualities with this method.

1 Introduction

Human speech carries large quantities of information. This can be measured objectively through physical parameters which can be matched to subjective, audible properties. For example, the physical parameter fundamental frequency (F0, frequency of vibration of the vocal folds during voiced speech) corresponds to the subjective pitch. In text-to-speech (TTS) synthesis, we can model and modify these physical properties in order to create speech that is similar to human speech.

State-of-the-art text-to-speech synthesis is based on statistical parametric methods. Particular attention is paid to Hidden Markov-model (HMM) based text-to-speech synthesis (Zen et al., 2007). Most TTS techniques can produce good quality and highly intelligible output. Recent studies showed, however, that current speech synthesis systems are still recognized as non-human when synthesizing extended passages (Keller, 2007; Németh et al., 2007). There are a number of ways to improve naturalness of the prosody of synthesized speech: van Santen et al. (2005) minimized the prosody modification artifacts in unit selection synthesis, Díaz and Banga (2006) combined the intonation modeling and speech unit selection, while Keller (2007) analyzed perceived rhythm to make synthesized speech less robotic. Identical or very similar pitch contours of successive sentences make the synthetic speech monotonous when synthesizing longer passages of text. The first goal of our work was to design a novel prosody module capable of generating more natural pitch contours and introducing variability over successive sentences.

The source-filter model of speech separates the source (glottal excitation) from the filter (traditionally the vocal tract) (Fant, 1960). This model has been successfully applied in various parts of speech technology (e.g. in speech coding). Recent research in speech synthesis used this model in the statistical parametric framework, in the HMM-based TTS. The speech signal was decomposed to source (excitation signal) and filter, the parameters of which are modeled separately and recombined during synthesis. The second goal of our work was to improve the way in which the source-filter model, particularly the excitation signal, is used in statistical parametric speech synthesis.

It was found that the source and filter of speech are not independent and non-linear interaction between source and filter may occur (Stevens, 1998; Titze, 2008). To model the source and filter properly, it is not enough to investigate the vocal tract, because the subglottal tract (the area below the glottis) has an influence on voice as well (Lulich, 2006). This area, the lower airways (consisting of the lungs, trachea and bronchi) has resonances similar to the formants of the vocal tract. These are called subglottal resonances (SGRs). Lulich (2010) found that SGRs have phonological effects in American English. Wang et al. (2009) found that subglottal resonances can cause formant attenuation and even jumps in the second formant curve. This can be applied in a field of speech technology: they have shown that SGRs are useful in speaker normalization. Our third goal was to investigate the effects of subglottal resonances on phonological distinctive features in Hungarian, and how these could be applied in speech synthesis. In the future, particular attention will be paid to investigate the role of subglottal resonances on the excitation signal.

The following sections are organized as follows. In Sec 2) we deal with the proper modeling of pitch variability in TTS systems. In Sec 3) we improve the excitation model used in speech synthesis and Sec 4) introduces research regarding subglottal resonances in human speech. Section 5) concludes the paper and shows how the above three topics may relate to each other.

2 Modelling prosodic variability in text-to-speech synthesis

There have been only a few studies regarding variable prosody in the TTS context. Chu et al. (2006) investigated the variation of prosody in human speech using a database containing two repetitions of 1000 recorded sentences in Mandarin. A synthesis approach to variable prosody has been addressed by Díaz et al. (2006) using a unit selection TTS system. This method preserved the intonation variability of the original speaker by selecting one of several pitch candidates.

In an initial study, we have experimented with introducing prosodic variability by F0 generation in a Hungarian diphone TTS environment (Németh et al., 2007). This method generates the F0 contour for a sentence to be synthesized based on a database of natural sample sentences. However, the similarity measure used between the input text and sentences from the database was not suitable for a general speech synthesizer.

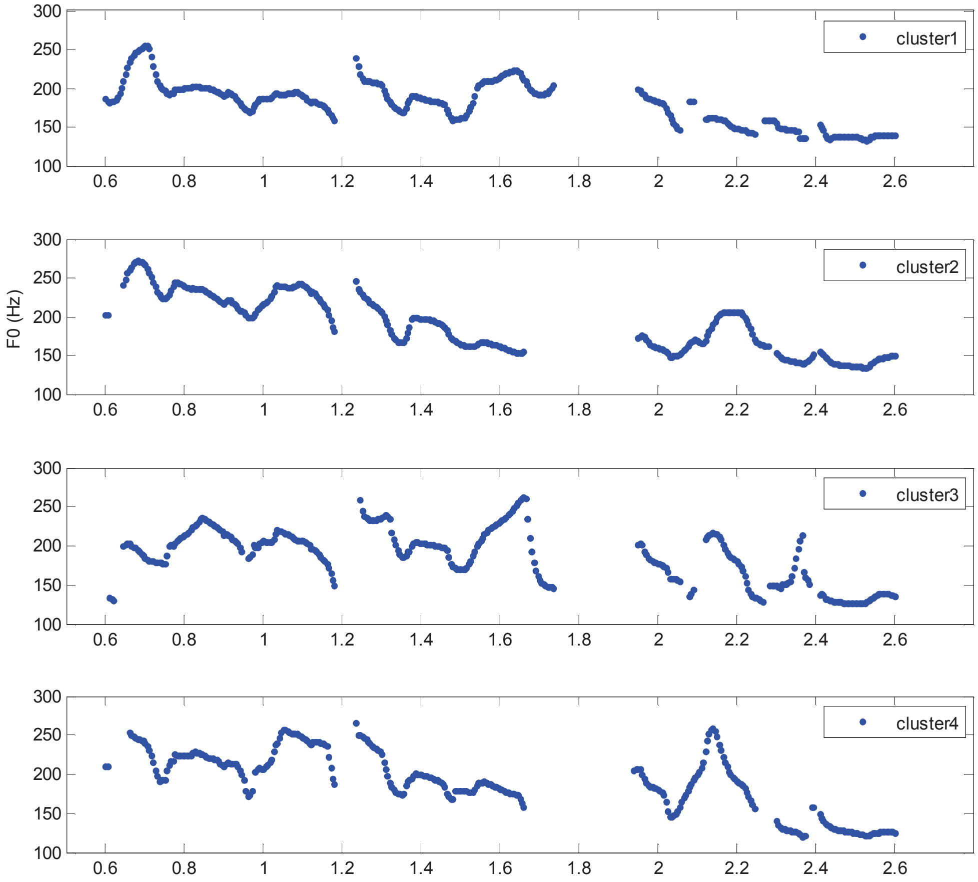

Figure 1. F0 contours of the four synthesized variants of the sentence *Zsigmond nem tagadja, hogy ő zsidó.* ('Sigismund does not deny that he is a Jew.')

Speech synthesis research has focused recently on statistical parametric methods; particularly HMM based speech synthesis (Zen et al., 2007). Here, the basic idea is that instead of hand-crafted rules for speech description, statistical machine learning methods are applied to analyze and synthesize speech. The parameters describing the speech signal are obtained via speech coding methods. These parameters are learned during HMM-based context clustering and assigned to the input text during synthesis. Finally, the speech is reconstructed from the parameters with a speech decoder.

HMM-based speech synthesis uses a training database of speech sentences with corresponding transcriptions. The parameters of speech (e.g., F0) are learned from this database. Typically, a few hours of speech from a single speaker is needed for acceptable quality. By splitting such a database to several subsets, and doing a separate HMM training on these subsets, the system learns different F0 models from each subset. In this way, we split a database to four subsets using the SOFM (Self-Organizing Feature Map, Kohonen et al. (1997)) unsupervised clustering method (Csapó and Németh 2011). We created four different F0 models for the HMM TTS system. During synthesis, a random F0 model is used, ensuring that the F0 contour of repeated synthesized sentences will likely be different.

Fig 1 shows an example for the results of our approach. The F0 contours of the four synthesized variants of a sentence are shown. We can see that the F0 contours are different (e.g., peaks are at different positions). Despite the differences in the F0 curve, the meaning of the four variants of the sentence is the same. Differences were audible in the pitch of the sentences according to a subjective listening test. Csapó and Németh (2011) contains an objective evaluation of this method, in which the F0 contour differences of four variants of 2000 sentences were measured.

3 Analysis and synthesis of the speech excitation signal

According to the source-filter theory, speech can be split into the source and filter (Fant, 1960). The source signal represents the glottal source that is created in the human glottis. The filter represents the vocal tract (including the mouth, tongue, lips, etc.). Traditionally, linear prediction coefficient (LPC) analysis can be used for the source-filter separation, but recently more complex and more accurate filtering methods have been used, including mel-spectrum and mel-generalized cepstrum (MGC) analysis (SPTK, 2011).

In the traditional HMM-based speech synthesis system, a very simple LPC vocoder is used for the source-filter model and an impulse sequence is used as the excitation in voiced parts, while unvoiced parts are modeled with white noise. However, this produces “buzzy” speech quality, for which HMM-based systems are often criticized. CELP (Codebook Excited LP) based methods offer the highest quality solutions to alleviate this problem (Drugman, 2011).

Fig. 2 shows the vocoding part within the HMM TTS framework. For the excitation, two types of signals are shown: 1) an impluse sequence and 2) the residual signal of speech that was obtained by MGC inverse filtering. The impulse sequence is an oversimplified model of the residual signal. The goal of our research was to synthesize the excitation signal that resembles the properties of real residual more properly than the impulse sequence.

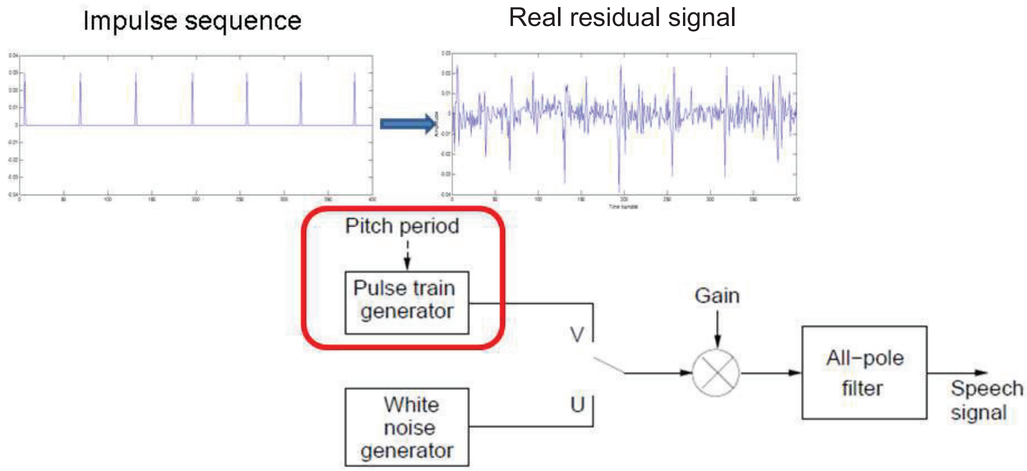

Figure 2. Vocoding within the HMM TTS framework. Two types of excitation signals are shown: the impulse sequence is an oversimplified model of the real residual signal.

Drugman (2011) was one of the the first to create such an excitation synthesis. He constructed a codebook of residual frames obtained from natural speech and used it in HMM synthesis. Cabral (2010) usesd the Liljencrants-Fant acoustic model of the glottal source derivative to construct the excitation signal. Raitio et al. (2011) used unit selection methods for the synthesis of excitation, where glottal periods obtained from real speech are concatenated resulting in a smooth excitation signal.

In our approach, we aimed to create a codebook-based excitation model that uses unit selection (Csapó and Németh, 2012). During the analysis part, the residual signal was obtained from natural speech with MGC-based inverse filtering. Starting from this signal, a codebook was built from phoneme-dependent pitch-synchronous excitation frames. Phoneme dependent frames were used, because we assumed that the inverse filtering was not perfectly decomposing source and filter, and that the residual signal may contain information regarding the phone as well. In other codebook-based vocoding methods, there is a general codebook obtained from all of the phones. Several parameters (e.g., period, T0, energy) of these frames are fed to the HMM training system. For the sentences to be synthesized, the HMM TTS assigns these parameters for each speech sound. During synthesis, phoneme-dependent excitation frames are selected from the codebook with unit selection, and concatenated to each other. After that, final synthesized speech is obtained with MGC-based filtering.

Subjective analysis of the quality that can be synthesized with this method has not been conducted yet, but our preliminary tests suggest that the quality is similar to other CELP based methods.

With this novel excitation approach, we will be able to model different voice qualities. By creating separate codebooks from breathy, whispered, or other type of speech, the method can synthesize speech with the specific voice quality.

4 Investigation of the relation between vowel formants and subglottal resonances

It has been shown for several languages that subglottal resonances play a role in dividing the frequency space of consonant and vowel acoustics into discrete regions corresponding to phonological categories. Lulich (2006) investigated American English speakers, Madsack et al. (2008) tested several speakers of two German dialects and Jung (2009) tested Korean speakers. SGRs have been reported to divide vowels into certain contrasting natural categories: low – non-low; front – back; front tense - lax. Our work aimed to consider the patterns of Hungarian vowels with regard to the SGRs in speech production and perception.

In a first experiment, we investigated the vowel space of four Hungarian speakers in nonsense word reading (Csapó et al., 2009). Subglottal resonances were measured from the accelerometer signal; the accelerometer was pressed to the neck while speaking. The results confirmed that the first subglottal resonance (Sg1) divides low and non-low vowels (in terms of the first formant, F1), while the second (Sg2) separates back and front vowels in Hungarian as well (in F2). The third subglottal resonance (Sg3) has been found to divide unrounded non-low front vowels from other front vowels. An example for this can be seen in Fig. 3. The figure shows the separating role of SGRs for a specific Hungarian speaker. The dividing line between low and non-low vowels (Sg1) is less clear compared to Sg2 and Sg3, possibly because it is more difficult to measure Sg1 than Sg2. The results are similar for other speakers as well (Csapó et al., 2009).

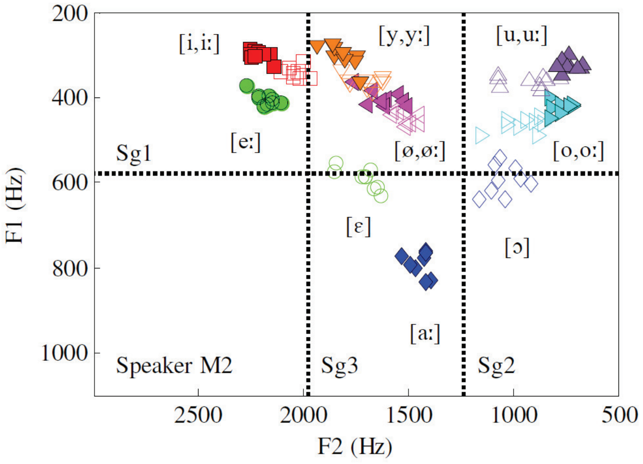

Figure 3. Formant space of a speaker with the 14 Hungarian vowels. The subglottal resonances are indicated by horizontal and vertical dashed lines.

In a second experiment, the formant spaces of six other speakers were analyzed in spontaneous speech (Csapó et al., 2011). We found that the SGR effects were less intensive in continuous speech compared to the nonsense word recordings. Particularly, data from most speakers did not support the role of Sg3.

In the next experiment, using data from the same six speakers, automatic formant-based vowel classification was applied and extended with normalization based on the subglottal resonances (Csapó et al., 2011). The input of the classification included the first two formant values of vowels extracted from spontaneous speech and the first three subglottal resonances. The target of the classification was groups corresponding to phonological distinctive features – e.g., in case of F2-Sg2 the target was 'back vs. front'. Three types of classifications were performed, of which two are shown here: 1) decision tree classification on raw formant data (without SGRs) and 2) decision tree classification on SGR-normalized formant data. Table 1 summarizes the correctly classified rates of these classifications. In 'low vs. non-low', Sg1 was used for F1 formant normalization, in 'back vs. front', Sg2 was used for F2 formant normalization and in 'front* vs. other', Sg3 was used for F2 formant normalization. According to the results of the experiment, it was shown that the knowledge of Sg2 and Sg3 may improve the accuracy of automatic vowel classification if using male and female data together in limited data contexts. Sg1 was not helpful in automatic classification of spontaneous speech.

Table 1: Result of the vowel classification experiments: correctly classified rates. Front* denotes the 'front unrounded non-low' vowel class.

	decision tree	**decision tree + SGR**
low vs. non-low	79.81%	78.09%
back vs. front	84.28%	84.73%
front* vs. other	86.95%	88.21%

Finally, in a pilot experiment, the perceived backness of the vowel [ɔ], as a function of F2, and Sg2 was investigated in a listening test (Csapó et al., 2011). C[ɔ]C transitions with different F2 frequencies at the vowel midpoint were extracted from two speakers' spontaneous recordings. Ordering the vowels by increasing F2, the results showed that for one of the two tested speakers, an abrupt increase in perceived backness of the vowel occurred when F2 was higher than Sg2. For the other speaker, a similar abrupt increase was not observed in the F2 – Sg2 relation.

It has been shown that SGRs can be applied in speaker normalization and automatic speech recognition (Wang et al., 2009; Arsikere et al., 2011). However, the usefulness of subglottal resonances in speech synthesis has been only initially investigated before. Gorbunov and Makarov (2011) modeled the subglottal region in an articulatory speech synthesizer by simulating the influence of the trachea, bronchi and and lungs. Hiroya (2011) introduced a method to remove the effect of subglottal resonances in speech signals for estimating a vocal-tract spectrum, and showed its accuracy in Japanese speech synthesis examples. It is possible that by modeling subglottal resonances explicitly in statistical parametric speech synthesis, the quality of this could be improved.

5 Summary

Speech synthesis has been tackled recently using statistical methods. In this study we have experimented with introducing prosodic variability by F0 generation in a Hungarian TTS environment. Our method generated several F0 contour variants for a sentence, and from this a random candidate could be chosen during synthesis. According to van Santen et al. (2005), this may be an important feature of future speech synthesis.

The HMM TTS technique makes use of large speech corpora, of which the main parameters of speech are extracted and later recombined. A significant problem of state-of-the art statistical parametric speech synthesis systems is their "buzzy" quality caused by an oversimplified excitation model. As part of my research, I am developing ways to more accurately model the excitation of a Hungarian statistical speech synthesis system, thereby improving the quality of synthetic speech. Compared to other excitation models (e.g. Drugman, 2011; Cabral, 2010; Raitio et al. 2011), my model is expected to produce similar quality synthesized speech. By further improving the excitation model, we will be able to synthesize different voice qualities (e.g. breathy, whispered) as well.

Investigating subglottal resonances in Hungarian helped to understand how speech production works. We have found relationships between vowel formants and SGRs that are similar to other languages (e.g. Lulich, 2006). It is not known whether the parameters currently used in HMM TTS properly model the phenomena caused by subglottal acoustics (e.g., jumps in the F2 track and formant attenuations) in synthesized speech.

The results regarding subglottal resonances have implications for understanding phonological distinctive features, as well as applications in automatic speech technologies. The latter includes speaker normalization (e.g. Wang et al., 2009) and other related problems in automatic speech recognition. The fact that SGRs are roughly constant for a given speaker may be useful in speaker recognition as well (Arsikere et al. 2011).

In the future, I plan to investigate the effect of SGRs on the excitation signal of speech obtained by inverse filtering. Several studies (on American English and Spanish) have confirmed that there are correlations between specific properties of the speech signal and SGRs, thus an indirect estimation of the resonances can be done using microphone recordings (Arsikere et al., 2011). Testing these algorithms on both Hungarian and English recordings will help to extend the hypothesis that subglottal resonances have relevance independent of the language used. A better understanding of how SGRs affect the glottal source will suggest new ways to improve the naturalness of synthesized speech.

In my PhD work I have been doing research in several subfields of speech science. In text-to-speech synthesis, our goal is to make the synthesized speech as close as possible to human speech. With my results, future speech synthesis is expected to become more natural.

Acknowledgements

I would like to thank the editor and the two anonymuous reviewers for the valuable suggestions and comments that have improved the manuscript. This research has been partially supported by the Paelife (Grant No. AAL-08-1-2011-0001), by the CESAR (Grant No. 271022) and by the TÁMOP-4.2.1/B-09/1/KMR-2010-0002 projects.

References

Arsikere, H., Lulich, S.M., and Alwan, A. 2011. Automatic estimation of the second subglottal resonance from natural speech. *ICASSP 2011*, 4616–4619.

Cabral, J. P., 2010. *HMM-based Speech Synthesis using an Acoustic Glottal Source Model.* PhD Thesis, CSTR, University of Edinburgh, United Kingdom.

Csapó, T.G., Bárkányi, Zs., Gráczi, T.E., Bőhm, T. and Lulich, S.M. 2009. Relation of formants and subglottal resonances in Hungarian vowels. *Interspeech 2009*, Brighton, United Kingdom, 484–487.

Csapó, T.G., Gráczi, T.E., Bárkányi, Zs., Beke, A. and Lulich, S.M. 2011. Patterns of Hungarian vowel production and perception with regard to subglottal resonances. *The Phonetician* 99-100, 7–28.

Csapó, T.G. and Németh, G. 2011. Prozódiai változatosság rejtett Markov-modell alapú szövegfelolvasóval, [Prosodic Variability in Hidden Markov Model-based Text-To-Speech Synthesis]. *MSZNY*, Szeged, Hungary, 167–177.

Csapó, T.G. and Németh, G. 2012. A novel codebook-based excitation model for use in speech synthesis. *CogInfoCom 2012*, Kosice, Slovakia, accepted.

Chu, M., Zhao, Y. and Chang, E. 2006. Modeling stylized invariance and local variability of prosody in text-to-speech synthesis. *Speech Communication* 48, 716–726.

Díaz, F.C. and Banga, E.R. 2006. A method for combining intonation modelling and speech unit selection in corpus-based speech synthesis systems. *Speech Communication* 48, 941–956.

Drugman, T. 2011. *Advances in Glottal Analysis and its Applications.* PhD Thesis, University of Mons, Belgium.

Fant, G. 1960. Acoustic theory of speech production. Mouton, The Hague.

Gorbunov, K.S. and Makarov, I.S. 2011.The subglottic region in articulator synthesizers. *Journal of Communications Technology and Electronics* 56, 1504–1509.

Hiroya, S., Miki, N., and Mochida, T. 2011. Multi-closure-interval linear prediction analysis based on phase equalization. *Proc. APSIPA*.

Jung, Y. 2009. *Acoustic articulatory evidence for quantal vowel categories: The features [low] and [back]*. PhD Thesis, MIT, USA.

Keller, E. 2007. Beats for individual timing variation. In A. Esposito et al. (eds.), *The Fundamentals of Verbal and Non-Verbal Communication and the Biometrical Issue*, IOS Press.

Kohonen, T., Kaski, S., and Lappalainen, H. 1997. Self-organized formation of various invariant-feature filters in the adaptive-subspace SOM. *Neural Computation*, vol. 9, no. 6, 1321–1344.

Lulich, S.M. 2006. *The Role of Lower Airway Resonances in Defining Vowel Feature Contrasts.* PhD Thesis, MIT, USA.

Lulich, S.M., 2010. Subglottal resonances and distinctive features. *Journal of Phonetics* 38(1), 20–32.

Madsack, A., Lulich, S. M., Wokurek, W., and Dogil, G. 2008. Subglottal resonances and vowel formant variability: A case study of high German monophthongs and Swabian diphthongs. *Proceedings of LabPhon 11*, 91–92.

Németh, G., Fék, M., and Csapó, T.G. 2007. Increasing Prosodic Variability of Text-To-Speech Synthesizers. *Interspeech 2007*, Antwerp, Belgium, 474–477.

Raitio, T., Suni, A., Yamagishi, J., Pulakka, H., Nurminen, J., Vainio, M., and Alku, P. 2011. HMM-Based Speech Synthesis Utilizing Glottal Inverse Filtering. *IEEE Transactions on Audio, Speech and Language Processing* 19(1): 153–165.

van Santen, J., Kain, A., Klabbers, E., Mishra, T. 2005. Synthesis of Prosody using Multi-level Unit Sequences, *Speech Communication*, Vol. 46(3–4), 365–375

SPTK working group. 2011. *Reference Manual for Speech Signal Processing Toolkit* Ver. 3.5, December 25, 2011.

Stevens, K.N. 1998. *Acoustic Phonetics*, MIT Press, Cambridge, MA.

Titze, I.R. 2008. Nonlinear source–filter coupling in phonation: Theory. *Journal of the Acoustical Society of America* 123, 2733–2749.

Wang, S., Lulich, S. M. and Alwan, A. 2009. Automatic detection of the second subglottal resonance and its application to speaker normalization. *Journal of the Acoustical Society of America* 126, 3268–3277.

Zen, H., Nose, T., Yamagishi, J., Sako, S. Masuko, T., Black, A.W. and Tokuda, K. 2007. *The HMM-based speech synthesis system version 2.0, ISCA SSW6.*