Investigating source and filter contributions, and their interaction, to statistical parametric speech synthesis

Thomas Merritt[1], *Tuomo Raitio*[2], *Simon King*[1]

[1]The Centre for Speech Technology Research, University of Edinburgh, U.K.
[2] Department of Signal Processing and Acoustics, Aalto University, Finland

t.merritt@ed.ac.uk, tuomo.raitio@aalto.fi, Simon.King@ed.ac.uk

Abstract

This paper presents an investigation of the separate perceptual degradations introduced by the modelling of source and filter features in statistical parametric speech synthesis. This is achieved using stimuli in which various permutations of natural, vocoded and modelled source and filter are combined, optionally with the addition of filter modifications (e.g. global variance or modulation spectrum scaling). We also examine the assumption of independence between source and filter parameters. Two complementary perceptual testing paradigms are adopted. In the first, we ask listeners to perform "same or different quality" judgements between pairs of stimuli from different configurations. In the second, we ask listeners to give an opinion score for individual stimuli. Combining the findings from these tests, we draw some conclusions regarding the relative contributions of source and filter to the currently rather limited naturalness of statistical parametric synthetic speech, and test whether current independence assumptions are justified.

Index Terms: speech synthesis, hidden Markov modelling, GlottHMM, source filter model, source filter interaction

1. Introduction

Despite great progress in statistical parametric speech synthesis in recent years, the naturalness of this type of synthetic speech is still too far below that of natural speech or the best unit selection systems [1, 2]. This is seen year after year in the results of the annual Blizzard Challenge [3–5, for example]. In the literature, explanations are proffered, but formal evidence is lacking. According to [1], the degradation stems mainly from three factors: a) over-simplified vocoder techniques that are incapable of representing natural speech waveforms in detail, b) acoustic modelling inaccuracy, and c) over-smoothing of the generated speech parameters.

However, these points are left somewhat vague and do not specify the exact causes of these effects, and there is no suggestion on what should be done to close the gap in naturalness (or 'quality'). The aim of this paper is to investigate points a) and c), and to discover which aspects of current systems are in most need of improvement.

Statistical parametric speech synthesis relies on the ability of the vocoder to decompose speech into a set of speech parameters that characterise the (perceptually) relevant aspects of speech. Most vocoders start from the source-filter model [6], where a speech waveform is modelled as a linear combination of an excitation signal and a resonant filter. In speech production, the corresponding components are the voice source and the vocal tract filter. In natural speech, the contributions of these components cannot be completely separated since they interact [7]. The existence of this interaction between source and filter is well known, yet rarely taken into account in speech technology. It is possible that the failure to model this interaction between the two components might be one cause of poor quality in statistical parametric speech synthesis.

Our aim here is to study the relative contributions of excitation and filter to the quality of synthetic speech, and to assess the degrading effect of vocoding and statistical modelling with regard to these two components, with a focus on the consequences of assuming that they are independent. Additionally, the effectiveness of three filter enhancement techniques is evaluated.

In order to do this, the GlottHMM vocoder is used in our experiments, because it is capable of decomposing speech into components corresponding closely to natural speech production: the glottal source signal and the vocal tract filter. A cross-synthesis scheme is adopted where speech is synthesised using the source and filter in all possible combinations of i) natural, ii) vocoded, and iii) modelled. The particular contribution of filter frequency response over-smoothing—or resonance sharpness—and changes in modulation characteristics are assessed by applying both enhancing and degrading effects to the filter. All these combinations were assessed in a large subjective evaluation where listeners rated the speech samples according to similarity to each other, and on a mean opinion score (MOS) scale.

2. Vocoder

As noted above, the GlottHMM vocoder [8] is used in the experiments, primarily because GlottHMM aims at the accurate modelling of the two speech production components: the voice source signal and the vocal tract filter. This type of vocoder, being closer to natural speech production than conventional vocoders (e.g. STRAIGHT [9, 10]), might be beneficial in testing hypotheses concerning source and filter contributions and interaction, and their effect in statistical speech synthesis. The other reason for choosing the GlottHMM vocoder is that it can be easily modified to accommodate the needs of this experiment; for example, it is straightforward to use a voice source signal derived from natural speech during synthesis. Finally since a previous study [11] was performed using STRAIGHT, it will be interesting to perform a comparable investigation with a different type of vocoder.

GlottHMM is based on the conventional source-filter model, but the decomposition of speech into two components is based on the physiology of human speech production: the voice source signal and vocal tract filter. GlottHMM uses iterative adaptive inverse filtering (IAIF) [12], a glottal inverse filtering method based on all-pole modelling for that purpose. After the decomposition, the voice source signal is parameterised in de-

Table 1: *Speech features extracted by the GlottHMM vocoder.*

Feature	Order	Source	Filter
Frame energy (dB)	1	×	
Log-fundamental frequency	1	×	
Harmonic-to-noise ratio (dB)	5	×	
Voice source spectrum LSF	10	×	
Vocal tract spectrum LSF	30		×

tail, in order to enable accurate reconstruction of the signal in synthesis. The speech features used by the GlottHMM vocoder are shown in Table 1. Moreover, GlottHMM uses a natural glottal flow waveform as a base for creating the synthetic excitation in order to preserve the phase characteristics of the natural glottal flow. GlottHMM has been shown to yield high-quality and very intelligible synthetic speech [8, 13–15], and it has already been used in various experiments investigating voice source modelling in statistical speech synthesis (e.g. [16–18]).

3. Experiments

3.1. Speech material and voice building

A speech database of a British male speaker was used in the study. The database consists of 2,022 read-aloud sentences selected for the purpose of speech synthesis, leading to approximately 2 hours of speech data (sampled at 16 kHz). The speech was parameterised using the GlottHMM vocoder, and an HMM-based voice was built following the standard HTS method [22]. Delta and delta-delta features were added to speech features, and semi hidden Markov models were used as acoustic models.

Table 2: *The 25 conditions investigated in the study, consisting of source and filter components from natural (nat), vocoded (voc), and modelled (hmm) speech. The filter processing methods are indicated in the last column (see definitions in Table 3).*

Condition name	Source	Filter	Filter processing
1-natural	nat	nat	
2-nat-voc	nat	voc	
2-nat-voc-ms−	nat	voc	MS–
2-nat-voc-smth	nat	voc	Smoothing
2-nat-voc-gv−	nat	voc	GV–
3-nat-hmm	nat	hmm	
3-nat-hmm-enh	nat	hmm	LSF-enh
3-nat-hmm-gv+	nat	hmm	GV+
3-nat-hmm-ms+	nat	hmm	MS+
4-voc-voc	voc	voc	
4-voc-voc-ms−	voc	voc	MS–
4-voc-voc-smth	voc	voc	Smoothing
4-voc-voc-gv−	voc	voc	GV–
5-voc-hmm	voc	hmm	
5-voc-hmm-enh	voc	hmm	LSF-enh
5-voc-hmm-gv+	voc	hmm	GV+
5-voc-hmm-ms+	voc	hmm	MS+
6-hmm-voc	hmm	voc	
6-hmm-voc-ms−	hmm	voc	MS–
6-hmm-voc-smth	hmm	voc	Smoothing
6-hmm-voc-gv−	hmm	voc	GV–
7-hmm-hmm	hmm	hmm	
7-hmm-hmm-enh	hmm	hmm	LSF-enh
7-hmm-hmm-gv+	hmm	hmm	GV+
7-hmm-hmm-ms+	hmm	hmm	MS+

All features were in individual streams except that the vocal tract spectrum LSFs and frame energy were in the same stream.

3.2. Cross-synthesis methodology

In order to study the relative effect of the source and filter components at each processing stage of speech synthesis, a cross-synthesis scheme is used where three versions (natural, vocoded, modelled), of each of the two components are created, from which all the permutations are used to synthesise speech. That is, we created stimuli that combined the properties of natural, vocoded and synthetic speech.

It is common in statistical speech synthesis to apply some enhancement to some speech parameters. The most common method is global variance (GV) [19]. We applied four different enhancement or degradation methods to the filter parameter trajectories: 1) global variance (GV) scaling [19], 2) scaling of modulation spectrum (MS) [20], 3) temporal smoothing as in [11], and 4) formant enhancement in the power spectrum domain [21]. One group of the stimuli starts from the vocoded filter and imposes certain properties of modelled speech by degrading the filter: the effect of statistical modelling is simulated by scaling the GV and MS of the LSF parameter trajectories to match the values seen in modelled speech, and also by temporal smoothing. Another group starts from a modelled filter and applies enhancement procedures aiming at improving the quality by scaling the GV up by 0.5 of the way from modelled towards natural, scaling the MS up by 0.85 of the way towards natural [20], and by applying formant enhancement to the LSFs in the power spectrum domain [21]. All the resulting 25 conditions resulting from the various combination are shown in Table 2 and the filter processing techniques are shown in Table 3.

In order to have a reference 'perfect' source, the GlottHMM vocoder was used to extract the voice source signal for the natural source conditions, given the filter for each condition. In order to combine the source and filter parts of each combination, the statistically modelled features were generated using time-aligned labels. To make sure the alignment between natural source/filter and synthetic source/filter were as good as possible, the voiced and unvoiced regions of vocoded and modelled fundamental frequency parameters were compared, and only the best matching sentences were used in the experiments.

3.3. Listening Tests

The perceptual testing for this investigation was in two phases, each employing a different paradigm: pairwise judgements analysed via multi-dimensional scaling (MDS), and mean opin-

Table 3: *The symbols and explanations for the processing methods applied to the filter parameter trajectories.*

Symbol	Explanation
GV– (for vocoded)	Global variance [19] scaled down to the level of synthetic speech
GV+ (for hmm)	Global variance [19] scaled up by 0.5 towards the level of natural speech
MS– (for vocoded)	Modulation spectrum [20] scaled down to the level of synthetic speech
MS+ (for hmm)	Modulation spectrum [20] scaled up by 0.85 towards the level of natural speech
Smoothing (for vocoded)	Smoothing the trajectory with a Hann window of length 21
LSF-enh (for hmm)	Formant enhancement applied to LSFs in the power spectral domain [21]

ion score (MOS) testing. The first of these involved listeners making "same or different quality" judgements about pairs of utterances generated under the differing conditions in Table 2. From these responses a perceptual distance matrix can be constructed, from which MDS generates a visualisation which plots the conditions in a multi-dimensional space. This method of testing has been found to be very illuminating when teasing apart the perceptual differences between speech stimuli [11,23], especially when it is suspected that listeners are using more than one perceptual dimension to make their judgements (something that a MOS test cannot discover). The second paradigm, MOS tests, required listeners to rate single stimuli (the same set of utterances as in the previous test) on a 5 point scale between 'bad' and 'excellent' in terms of the quality of the speech. Whilst MDS is potentially quite powerful, it can sometimes be difficult to draw precise conclusions from the complex plots is produces; the MOS test was selected to provide a basis for the interpretation of the MDS analysis.

3.3.1. Pairwise listening test

In the first listening test, listeners were presented with pairs of stimuli in which every condition was paired with every other condition (but not itself). Each possible pair of 25 conditions was repeated 12 times resulting in $12 \times (25^2 - 25) = 7200$ pairs. The pairs were presented in a randomised order to minimise bias. The sentence (i.e., text) was different for each utterance within a pair, with the sentences being drawn otherwise at random from a set of 40 sentences. The presentation order of the pairs was such that no sequence of two pairs contained the same sentence more than once, but was otherwise random. The 7200 pairwise comparisons were divided amongst 45 listeners, with each listener making 160 pairwise quality judgements. This number of judgements has previously been demonstrated to be reasonable for subjects [23].

3.3.2. Single stimulus listening test

In the MOS test, one utterance for each of the testing conditions (selected at random per condition) was presented to each listener 4 times. Thus, each listener evaluated 100 samples. The presentation order was such that no sequence of two utterances involved either the same sentence or the same condition, but was otherwise random.

4. Results

4.1. MDS plot

MDS generates a visualisation in a specified number of dimensions. The MDS stress factor [23] is an indication of whether the visualisation is an accurate representation of the distance matrix. The visualisation is therefore a trade off between a low dimensionality, which makes interpretation easier, and a fair representation of the listener responses. The stress factor indicated that 2 dimensions gives a reasonable representation of our listener responses.

In the visualisation plot, there is one point for each condition. Those conditions which listeners judged to be more perceptually similar will be closer together in the plot. Fig. 1 shows the MDS plot at 2 dimensions, about which we make the following observations:

Voice source The points (i.e., conditions in Table 2) appear to cluster in 4 groups: A) natural speech (*1-natural*) & natural source and perfectly matched filter without modifications

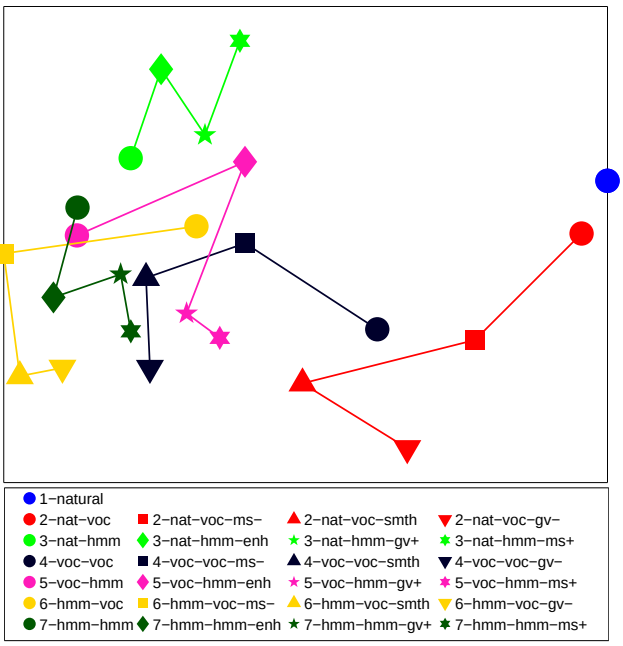

Figure 1: *MDS plot for 2 dimensions.*

(*2-nat-voc*); B) natural source and vocoded filter with degradations applied (*2-nat-voc-x*) & pure vocoded speech (*3-voc-voc*); C) natural source and modelled filter (*3-nat-hmm*) & vocoded source and modelled filter with formant enhancement (*5-voc-hmm-enh*); D) all other conditions. This shows that using the natural source is the biggest single factor, indicating that a better source signal has the potential to substantially improve the quality of the resulting speech. However, the arrangement of the clusters also tells us that any mismatch between source and filter has damaging perceptual consequences.

Vocoded and modelled voice sources (*5-voc-hmm* and *7-hmm-hmm*) are very close to one another when using a modelled filter, indicating that the modelling of the source is not a restricting factor in this situation. However, the filter enhancements are slightly more effective in the case of vocoded source than when combined with the modelled source.

Vocal tract filter Using a vocoded source in combination with a vocoder filter (*4-voc-voc*) or modelled filter (*5-voc-hmm*) are perceptually very similar in the cases when the vocoder filter is degraded, and the modelled filter is enhanced. This suggests that these enhancements are working when applied to modelled filters, although they do not quite make the speech the same as vocoded speech: listeners can still easily distinguish them.

The perceptual closeness of conditions with vocoded source and modelled filter (*5-voc-hmm*) and HMM synthesis (*7-hmm-hmm*) would strongly suggest that the current quality of statistical parametric speech synthesis systems (that make a source / filter independence assumption), is limited mainly by the modelling of the filter. The *6-hmm-voc* condition is closer to natural speech than either of these two conditions, further supporting this conclusion.

Source and filter interaction The perceptual distance between the conditions with modelled source and vocoded filter (*6-hmm-voc*) and the HMM synthesis conditions (*7-hmm-hmm*) are generally small, once degradation and enhancement effects are applied respectively. This is interesting: applying the enhancement to HMM-synthesis appears *not* to make the speech quality noticeably different, in contrast to vocoded source and

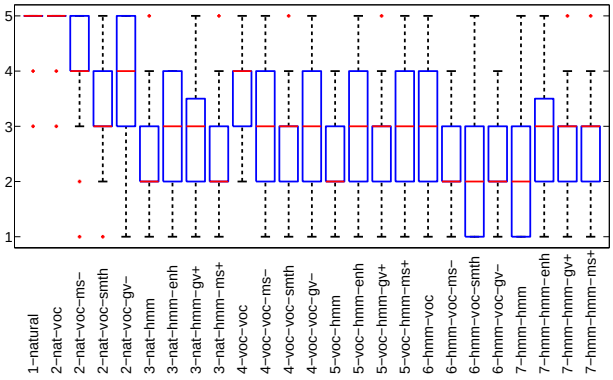

Figure 2: *Box plot of listener opinion scores*

modelled filter (*5-voc-hmm*). This could indicate either: 1) there is something natural about the vocoded source that modelling fails to capture; or 2) that once the source and filter have been independently modelled, nothing much can be done to recover from that.

The large perceptual distance between systems with natural source and vocoded filter (*2-nat-voc*) and natural source and modelled filter (*3-nat-hmm*) (among the largest between system configurations) can be interpreted in two ways: 1) it could be caused by artefacts introduced by mismatches between source and filter in *3-nat-hmm*; or 2) it could be due to the differences between vocoded and modelled filter coefficients when excited by the 'perfect' natural source, resulting in a match between source and filter in one condition and not in the other.

Natural source with vocoded filter conditions using MS degradation *2-nat-voc-ms−* and smoothing *2-nat-voc-smth* both lie perceptually close to vocoded speech *4-voc-voc*. This may indicate that MS down-scaling and smoothing both introduce mismatch between the source and filter, similar to the effect introduced by vocoding with current source-filter models.

4.2. MOS scores

The results of the MOS test are shown in Fig. 2. These largely back up the main conclusions from the MDS analysis: that distance from the natural speech point in the MDS plot corresponds closely to decrease in speech quality, and that the filter enhancements applied to the HMM filter produce noticeable improvements in the quality of speech.

An interesting contradiction between the results from the MDS and MOS tests is the scores for the natural source with modelled filter configuration (*3-nat-hmm*). In the MDS plot, the conditions closest to natural speech were GV and MS up-scaling (GV+ and MS+). However the results from the opinion score test showed that listeners prefer the speech with formant enhancement (LSF-enh) and GV up-scaling (GV+). This shows that listeners in the MDS test were not simply making one-dimensional preference comparisons and were instead making their judgements along more than one dimension of difference.

Other points of interest from these results include:

Voice source Conditions *2-nat-voc-ms−* and *4-voc-voc* receive similar quality scores, which coincides with the findings of the MDS test that these are perceptually similar. However *2-nat-voc-smth* is not rated as highly and instead there is a preference for *2-nat-voc-gv−*, which was the furthest point in the *2-nat-voc* system configuration in the MDS plot, highlighting that speech produced under this set of conditions is high in quality.

Vocal tract filter The MOS results for vocoded speech (*4-voc-voc*) and for vocoded source and modelled filter (*5-voc-hmm*) support the findings of the MDS test, in that speech under these conditions, following degradations and enhancements respectively, have very similar quality. This supports our observation that the effects caused by statistical modelling are being perceptually repaired to some extent, but that the speech is still of noticeably worse quality than vocoded speech.

The modelling of the filter parameters in *5-voc-hmm* may be a key factor limiting the quality output when source and filter are determined independently, as the quality score of *5-voc-hmm* and *7-hmm-hmm* remain similar whereas *6-hmm-voc* is rated better in quality by listeners. However this test found little difference between *6-hmm-voc* and the *5-voc-hmm* configurations once filter enhancements are applied.

Source and filter interaction The results for natural source and vocoded filter show the largest quality drop when using smoothing (*2-nat-voc-smth*). Smoothing of speech trajectories may create the largest decrease in the interaction, or degree of consistency, between the source and filter, by averaging content across consecutive frames and removing all fine variations from the trajectories. MS and GV degradation have less effect than smoothing, presumably because they are preserving more of the source-filter interaction, in other words, that the frame-by-frame variations in the filter parameters are consistent with the frame-by-frame variations in the source.

Applying enhancements to HMM-based speech (*7-hmm-hmm*) does not help as much as when they are applied to the condition using a vocoded source and HMM filter (*5-voc-hmm*). Possible explanations were already offered for this in Sec. 4.1.

5. Conclusion

A framework has been presented which makes it possible to investigate the effects introduced by the modelling of source and filter coefficients and the effectiveness of three filter enhancement techniques. By creating appropriate stimuli, performing two listening tests, and analysing the results, it has been possible to see clear differences in quality as source and/or filter are varied from natural, through vocoded to HMM modelled.

Current filter enhancement techniques are able to recover some of the quality loss caused by modelling the filter, yet the final quality seems to be more affected by the interaction of source and filter than by the individual quality of either one alone. Whilst it is impossible to 'prove' anything beyond reasonable doubt using perceptual tests, we are nevertheless able to say that the assumption of independence between source and filter, which is inherent in all current statistical parametric speech synthesisers, is one of the most significant limiting factors on the quality of synthetic speech.

The next step is to more carefully investigate the assumptions and findings made in this study. Prospective work includes assessing the importance of the synchronisation of the excitation instants and the filter parameters, and research into new models that effectively capture the interaction between the source and filter.

6. Acknowledgements

This work was supported by EPSRC Programme Grant EP/I031022/1 (Natural Speech Technology) and by the European Community's Seventh Framework Programme (FP7/2007-13) project under grant agreement No. 287678 (Simple4All).

7. References

[1] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Commun.*, vol. 51, no. 11, pp. 1039–1064, 2009.

[2] S. King, "An introduction to statistical parametric speech synthesis," *Sadhana*, vol. 36, no. 5, pp. 837–852, 2011.

[3] S. King and V. Karaiskos, "The Blizzard Challenge 2009," in *Proc. Blizzard Challenge*, Edinburgh, United Kingdom, 2009.

[4] ——, "The Blizzard Challenge 2010," in *Proc. Blizzard Challenge*, Kansai Science City, Japan, 2010.

[5] ——, "The Blizzard Challenge 2011," in *Proc. Blizzard Challenge*, Turin, Italy, 2011, pp. 185–190.

[6] G. Fant, *Acoustic Theory of Speech Production*. The Hague: Mouton, 1960.

[7] I. R. Titze, "Nonlinear source-filter coupling in phonation: Theory," *J. Acoust. Soc. Am.*, vol. 123, no. 5, pp. 2733–2749, 2008.

[8] T. Raitio, A. Suni, J. Yamagishi, H. Pulakka, J. Nurminen, M. Vainio, and P. Alku, "HMM-based speech synthesis utilizing glottal inverse filtering," *IEEE Trans. on Audio, Speech, and Lang. Proc.*, vol. 19, no. 1, pp. 153–165, 2011.

[9] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigné, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: possible role of a repetitive structure in sounds," *Speech Commun.*, vol. 27, no. 3–4, pp. 187–207, 1999.

[10] H. Kawahara, J. Estill, and O. Fujimura, "Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system STRAIGHT," in *2nd International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications (MAVEBA)*, 2001.

[11] T. Merritt and S. King, "Investigating the shortcomings of HMM synthesis," in *8th ISCA Workshop on Speech Synthesis (SSW8)*, pp. 185–190.

[12] P. Alku, "Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering," *Speech Commun.*, vol. 11, no. 2–3, pp. 109–118, 1992.

[13] A. Suni, T. Raitio, M. Vainio, and P. Alku, "The GlottHMM speech synthesis entry for Blizzard Challenge 2010," in *Proc. Blizzard Challenge 2010 workshop*, 2010, http://festvox.org/blizzard.

[14] ——, "The GlottHMM entry for Blizzard Challenge 2011: Utilizing source unit selection in hmm-based speech synthesis for improved excitation generation," in *Proc. Blizzard Challenge 2011 workshop*, 2011, http://festvox.org/blizzard.

[15] ——, "The GlottHMM entry for Blizzard Challenge 2012 - hybrid approach," in *Proc. Blizzard Challenge 2012 workshop*, 2012, http://festvox.org/blizzard.

[16] T. Raitio, A. Suni, H. Pulakka, M. Vainio, and P. Alku, "Utilizing glottal source pulse library for generating improved excitation signal for HMM-based speech synthesis," in *Proc. IEEE Int. Conf. on Acoust. Speech and Signal Proc. (ICASSP)*, 2011, pp. 4564–4567.

[17] T. Raitio, A. Suni, M. Vainio, and P. Alku, "Synthesis and perception of breathy, normal, and lombard speech in the presence of noise," *Computer Speech & Language*, vol. 28, no. 2, pp. 648–664, 2014.

[18] ——, "Comparing glottal-flow-excited statistical parametric speech synthesis methods," in *Proc. IEEE Int. Conf. on Acoust. Speech and Signal Proc. (ICASSP)*, 2013, pp. 7830–7834.

[19] T. Toda and K. Tokuda, "A speech parameter generation algorithm considering global variance for HMM-based speech synthesis," *IEICE Trans. Inf. Syst.*, vol. E90-D, no. 5, pp. 816–824, 2007.

[20] S. Takamichi, T. Toda, G. Neubig, S. Sakti, and S. Nakamura, "A postfilter to modify the modulation spectrum in HMM-based speech synthesis," in *Proc. IEEE Int. Conf. on Acoust. Speech and Signal Proc. (ICASSP)*, May 2014.

[21] T. Raitio, A. Suni, H. Pulakka, M. Vainio, and P. Alku, "Comparison of formant enhancement methods for HMM-based speech synthesis," in *7th ISCA Workshop on Speech Synthesis (SSW7)*, Sep. 2010, pp. 334–339.

[22] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A. Black, and K. Tokuda, "The HMM-based speech synthesis system (HTS) version 2.0," in *6th ISCA Workshop on Speech Synthesis (SSW6)*, Aug. 2007, pp. 294–299.

[23] C. Mayo, R. A. Clark, and S. King, "Listeners weighting of acoustic cues to synthetic speech naturalness: A multidimensional scaling analysis," *Speech Communication*, vol. 53, no. 3, pp. 311–326, 2011.