ATTRIBUTING MODELLING ERRORS IN HMM SYNTHESIS BY STEPPING GRADUALLY FROM NATURAL TO MODELLED SPEECH

Thomas Merritt[1]*, Javier Latorre*[2]*, Simon King*[1]

[1] The Centre for Speech Technology Research, University of Edinburgh, UK.
[2] Toshiba Research Europe Ltd., Cambridge Research Lab, Cambridge, UK.

T.Merritt@ed.ac.uk, javier.latorre@crl.toshiba.co.uk, Simon.King@ed.ac.uk

ABSTRACT

Even the best statistical parametric speech synthesis systems do not achieve the naturalness of good unit selection. We investigated possible causes of this. By constructing speech signals that lie inbetween natural speech and the output from a complete HMM synthesis system, we investigated various effects of modelling. We manipulated the temporal smoothness and the variance of the spectral parameters to create stimuli, then presented these to listeners alongside natural and vocoded speech, as well as output from a full HMM-based text-to-speech system and from an idealised 'pseudo-HMM'. All speech signals, except the natural waveform, were created using vocoders employing one of two popular spectral parameterisations: Mel-Cepstra or Mel-Line Spectral Pairs. Listeners made 'same or different' pairwise judgements, from which we generated a perceptual map using Multidimensional Scaling. We draw conclusions about which aspects of HMM synthesis are limiting the naturalness of the synthetic speech.

Index Terms— speech synthesis, hidden Markov modelling, vocoding

1. INTRODUCTION

HMM synthesis remains significantly behind the quality of natural speech and speech output from concatenative (unit selection) synthesis systems under 'best-case' conditions, as repeatedly highlighted in the results from Blizzard Challenges [1, 2, 3, for example], even though much progress has been made [4, 5]. Whilst the HMM approach is relatively robust when it comes to handling training data with poor phonetic coverage or low recording quality [6], it fails to produce natural-sounding speech even when plentiful high-quality data are available [7].

Various explanations have been postulated regarding the cause of this apparent ceiling effect, the most common including: over-smoothing of the spectral envelope as a consequence of averaging over multiple speech samples [8, 9]; over-smoothing of the parameter trajectories due to the MLPG algorithm [10, 11]; poor performance of vocoders [12], particularly regarding source-filter separation. However, these theories are only occasionally tested in formal studies [13, 14, 15].

Following a methodology that we proposed earlier [13], the current study adds a number of novel contributions: the use of idealised 'pseudo-HMMs' which only involve averaging a few contiguous frames from a single training example aligned with one HMM state, and so remove the effect of across-class averaging (explained in Section 3.2); two different speech parameterisations; the use of a commercial-quality speech database; the inclusion of natural (not vocoded) waveforms; the inclusion of a complete text-to-speech system. Our initial study [13] was limited to adjusting temporal smoothness and variance of the speech parameters in vocoded speech. As in [13], here we also focus on the spectral parameters. In all stimuli presented to listeners, the original, natural phone durations (found using forced alignment) and F0 were used. In related studies we have also investigated the relative contributions of source and filter [14] and the independence assumptions made by the statistical models [15].

Condition	Speech signal origin	Hanning smoothing window duration (frames)	Standard deviation scaling (%)
hann-1-stddev-080	vocoded	none	80
hann-5-stddev-080	vocoded	5	80
hann-11-stddev-080	vocoded	11	80
hann-21-stddev-080	vocoded	21	80
Vocoded	vocoded	none	100
hann-5-stddev-100	vocoded	5	100
hann-11-stddev-100	vocoded	11	100
hann-21-stddev-100	vocoded	21	100
hann-1-stddev-120	vocoded	none	120
hann-5-stddev-120	vocoded	5	120
hann-11-stddev-120	vocoded	11	120
hann-21-stddev-120	vocoded	21	120
hann-1-stddev-140	vocoded	none	140
hann-5-stddev-140	vocoded	5	140
hann-11-stddev-140	vocoded	11	140
hann-21-stddev-140	vocoded	21	140
hann-5-stddev-match	vocoded	5	match original
hann-11-stddev-match	vocoded	11	match original
hann-21-stddev-match	vocoded	21	match original
HMM-synth	HMM (with GV)	none	100
Original	natural	N/A	N/A
pseudo-HMM	pseudo-HMM	none	100

Table 1. *The 22 conditions presented to listeners*

2. METHODOLOGY

The framework introduced in [13] simulates various effects of modelling speech parameters in an HMM framework. Whilst the approach is general and extensible in principal, it was only used to in-

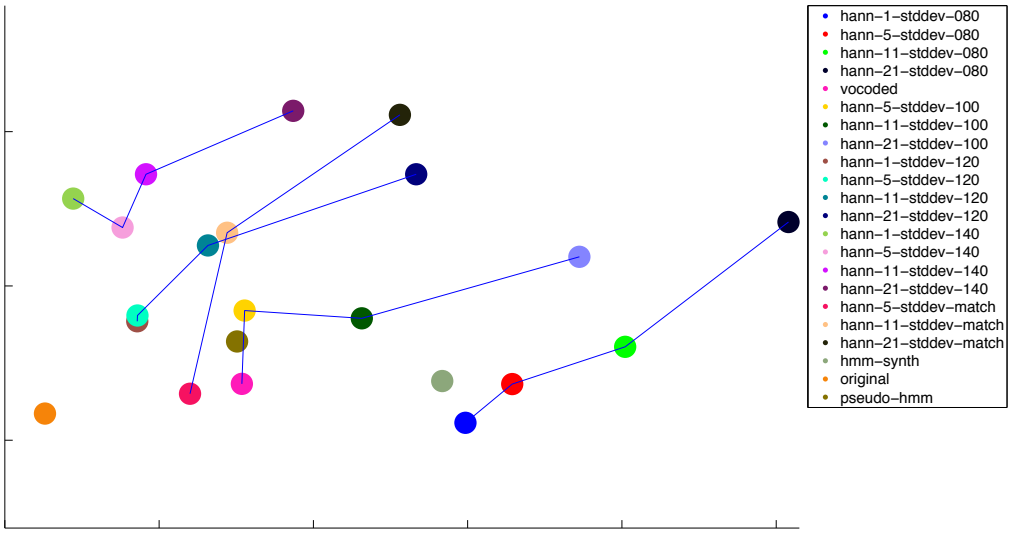


Fig. 1. *X-Y projection of the Mel-Cepstral MDS space. Lines have been added, connecting points with the same amount of variance modification but differing amounts of smoothing. The point for natural speech is in the lower left corner. We can infer that points closer to this correspond to more natural-sounding speech.*

vestigate the perceptual effects of temporal smoothing and incorrect (too large or too small) variance in the trajectories of speech spectral envelope (i.e., filter) parameters. Here, we follow the methodology laid out in [13] – and we refer the reader there for a full description – but go substantially further in terms of the modelling effects that we investigate. In brief, the methodology involves creation of various stimuli through simulations of HMM modelling effects, a pairwise "same or different quality" listening test, and analysis of the responses using multidimensional scaling (MDS) which provides a visualisation of the stimuli in which distance corresponds to perceived degree of difference, and in which it is possible to see whether listeners use more than one dimension when judging that difference.

3. CREATING THE SPEECH STIMULI

All natural and vocoded speech samples were based on speech from a male speaker (*mgt*) from the Toshiba *Studio-HQ* database [16]. This is a professional speaker recorded in a high quality studio, speaking in a neutral style. 1456 sentences from the same speaker were used to train the models.

3.1. Speech parameters

Spectral parameters were extracted with a Fourier transform using pitch synchronous windowing, then transformed into Mel-Cepstral or Mel-LSP coefficients [17] using SPTK [18]. The aperiodic energy was estimated using a pitch-scaled harmonic filter [19] and parameterized into 23 bark-scaled aperiodicity bands.

3.2. Simulating the effects of modelling

The standard approach to statistical parametric speech synthesis uses HMMs with a fixed number of emitting states [6], each containing a multivariate Gaussian distribution. When generating from such a model using the MLPG algorithm [10], a sequence of frames is emitted from each state: the mean of those frames is constant over the duration of the state. This introduces an effect of *temporal smoothing* over the generated parameters, and the amount of smoothing varies with the state duration.

The state means are estimated from data by averaging (typically via Expectation-Maximisation) the speech parameters from the contiguous sequence of frames associated with that state. This introduces *within-class spectral smoothing*. Furthermore, since no training database can include sufficient examples of every class (i.e., phonetic-prosodic context-dependent phoneme), examples drawn from differing contexts must be pooled and averaged together in order to robustly estimate the state mean and variance. This introduces further *between-class spectral smoothing*.

In addition to this, the *variance* of the generated trajectories may not match that of natural speech, due to the estimation of the model parameters from limited data, and/or inadequate models, and/or the parameter generation method.

Temporal smoothing: This effect was implemented exactly as in [13], to simulate the temporal smoothness of speech parameters generated by MLPG.

Variance adjustment: Again, this was implemented exactly as in [13], to simulate the potentially-incorrect variance of those trajectories (which can occur even when the GV technique [8] is employed). Previous investigations [20] have found that this method can enhance the speech as much as GV.

Parameter averaging: We hypothesised that the effect of *within-class* averaging over short sequences of contiguous frames from a single training example is small compared to *between-class* averaging of frames drawn from differing contexts. To test this, we constructed an idealised "pseudo-HMM" in which within-example averaging is present, but there is no between-class averaging. For comparison, we also used a complete, speaker-dependent HMM system similar to that described in [21], which of course does involve both within-class and between-class averaging across frames drawn from different contexts. The pseudo-HMM is created by using a natural example of the sentence to be 'synthesised', to ensure that the contexts are an exact match. For each such individual utterance, an association between states and frames was obtained by forced alignment using a speaker-dependent HMM. The mean value of each state was

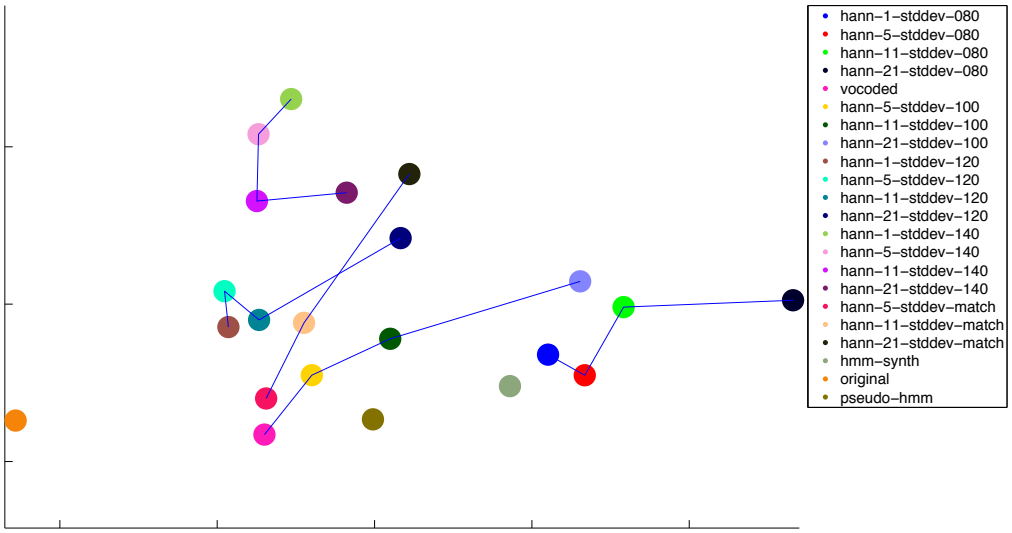


Fig. 2. *The Mel-LSP MDS space. Lines have been added to aid readability, as in Figure 1.*

computed as the median[1] of the frames associated with that state.

During the synthesis with either HMM and pseudo-HMM, the phone and state durations of the corresponding natural utterance were copied. For the excitation signal, the original F0 was first made continuous by interpolating through unvoiced regions, then combined with the aperiodicity values to generate mixed-excitation [22]. In the case of the HMM, the aperiodicity values were those generated by the model, to ensure consistency with the generated spectral envelope[2].

The manipulations performed by temporal smoothing, variance adjustment, and the pseudo-HMM may create inconsistent speech parameters in the Mel-LSP case: it is possible for the coefficient trajectories to become too close together, to cross, or to exceed the Nyquist frequency. Therefore, in order to reduce consequent artefacts, Mel-LSP values were limited to the interval $(0, \pi)$, whilst maintaining a reasonable spacing between coefficients and with the limits of the interval (we used an arbitrary value of 0.01 rad). Wherever two coefficients crossed (i.e., were not in ascending order), the lower coefficient was reduced so that it was at most 0.01 rad lower than the higher-order coefficient[3]. These corrections successfully removed all audible artefacts.

4. EXPERIMENTS

The strengths of the temporal smoothing and variance adjustment modifications were chosen by informal listening, so that they created a similar range of imperfections to those found in the speech from the *HMM-synth* condition (Table 1). The values chosen were scaling the standard deviation by 80, 100 (i.e., no modification), 120 and 140% as well as a scaling value (*stddev-match*) chosen such that the final global standard deviation matched that measured before application of temporal smoothing. The full range of conditions presented to listeners for pairwise comparison is shown in Table 1.

[1]Median was used instead of mean because it is more robust when the number of frames is small, which is the case here.

[2]Aperiodicity and spectrum are strongly related. Even with continuous F0, if a voiced spectrum is mixed with an unvoiced aperiodicity the result is a harsh noise. To avoid such mismatch affecting the judgments, synthetic aperiodicity was used with synthetic spectrum.

[3]In work to be published shortly, we have found this method performs better than SPTK's lspcheck.

In the listening test, listeners were asked to make forced-choice 'same or different quality' judgements about pairs of stimuli. 30 held-out test sentences taken from the same speaker used to train the models (Section 3) were used for testing, to which each of the 22 selected conditions in Table 1 were applied. The two items in each pair were differing sentences (randomly selected from the 30) processed under differing conditions (all possible pairs of conditions were covered). Every pair of stimuli was presented a total of 15 times. This resulted in a grand total of $15(22^2 - 22) = 6930$ pairwise comparisons. Each of the 45 listeners in the test was asked to make 154 'same or different quality' judgements, selected randomly without replacement from the 6930 pairs. This number of judgements is within the limit which an individual listener can tolerate [23].

The entire listening test was run twice: once using stimuli constructed using Mel-cepstra parameters, then using Mel-LSP. In both listening tests, the *Original* speech waveform was also included. The outcome of each test is a matrix in which each cell contains the number of 'different' judgements, summed across listeners: i.e., a matrix of perceptual distances. Multidimensional scaling (MDS) [24] is then used to visualise this matrix of distances, where each condition is a point in multi-dimensional space and distances between points reflect the perceptual distances from the matrix. We used the Matlab implementation[4] of MDS with Kruskal's normalised STRESS1 criterion [25].

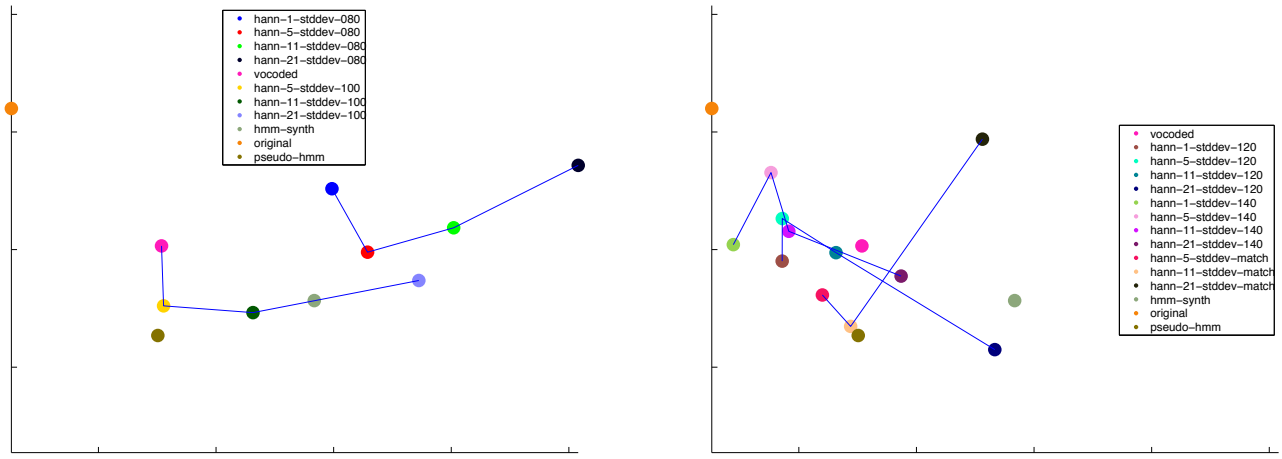
5. RESULTS

5.1. Mel-cepstral parameterisation

The stress levels (not reported here for reasons of space) suggested that 3 dimensions gave a reasonable representation of the perceptual space for this listening test. Here we examine this space, 2 dimensions at a time.

Figure 1 plots the 2-dimensional X-Y projection of the 3-dimensional MDS space. Scaling the standard deviation of the speech parameters appears to correspond to a lower-right to upper-left movement. As variance is scaled from 80% to 140%, the speech becomes first closer to natural (at 100% and 120%) and eventually moves further away, as we would expect. The *hann-5-stddev-match* condition, which is the same as *vocoding* (which we could also

[4]mdscale from the Matlab statistics toolbox



(a) *Reduced variance data points, plus references.*

(b) *Increased variance data points, plus references.*

Fig. 3. *X-Z projection of the Mel-cepstral MDS space.*

denote *hann-1-stddev-match*) but with very light smoothing applied, comes approximately as close to natural speech as vocoded speech does. This suggests that removing the fine temporal detail in speech parameter trajectories is not detrimental : it is probably noise arising from parameter estimation, not speech information. Excessive smoothing (*hann-21-...*) pushes the speech quality far away from natural.

Figure 3 shows the X-Z projection of the MDS space, with the reduced and increased variance conditions plotted separately, for clarity. In Figure 3a the points move towards HMM speech as a small amount of smoothing is applied and then move away again. Figure 3b reveals that applying increasing amounts of smoothing to the increased variance conditions initially gets us closer to natural speech before then moving away.

The pseudo-HMM condition is fairly close to conditions with variance unscaled (100%) or matched to the vocoded speech, with light or moderate smoothing. It is also considerably closer to vocoded speech (which is an upper bound for the pseudo-HMM) than the true HMM condition (*hmm-synth*). Together, these suggest that between-class averaging is indeed more harmful to naturalness than within-class averaging.

Conclusions: Matching the variance of natural speech is important. Whilst too much or too little variance both sound less natural, it is probably better to have slightly too much variance than too little. Light smoothing appears not to be detrimental, presumably because it merely removes minor artefacts created during parameter extraction from the original speech signal. Apart from getting the variance in the right range (equal to or slightly greater than that of vocoded speech), between-class averaging is the single biggest cause of reduced naturalness of HMM synthetic speech.

5.2. Mel-LSP parameterisation

The stress factors suggested that 2 dimensions gave a fair representation of the perceptual space in this case. Figure 2 shows that conditions with slightly increased variance (standard deviation scaled to 120%), conditions with variance matching that of vocoded speech, and vocoded speech itself, are all perceptually about the same distance from natural speech. However, excessive variance (standard deviation scaled to 140%) are highly detrimental. As with the Mel-Cepstral case, light smoothing does no harm, but heavy smoothing causes large reductions in naturalness.

Reducing the variance of the Mel-LSP parameters (standard deviation scaled to 80%) quickly moves the speech a large perceptual distance away from natural and vocoded speech. The HMM speech (*hmm-synth*) lies very close to some of the conditions with reduced variance (standard deviation scaled to 80%) and light to moderate smoothing, suggesting that the HMMs (despite the use of GV) fail to generate speech parameters with adequate variance. The pseudo-HMM condition is considerably closer to vocoded speech than the true HMM condition (*hmm-synth*), which suggests again that between-class averaging is indeed harmful to naturalness.

Conclusions: In the case of Mel-LSP parameterisation, too little variance is highly damaging and it is clearly better to have slightly more variance (than vocoded speech). Light smoothing is not problematic, but neither is it beneficial. As with the Mel-cepstral parameterisation, averaging the contiguous sequence of frames aligned with a single HMM state (*pseudo-hmm*) degrades the speech a little, but not as much as averaging across different contexts (*hmm-synth*).

6. SUMMARY

The simulation framework we introduced in [13] has been used to compare a much wider range of conditions. Our experimental results lead us to draw these conclusions:

- generating speech parameters with the correct variance is preferred, as in [13], but we now add that erring on the side of slightly too much variance is much better than too little;
- small amounts of temporal smoothing are not harmful, as in [13].
- within-class averaging of short contiguous sequences of frames is mildly harmful, in the same way as excessive smoothing;
- across-class averaging is very harmful.

Acknowledgements: this work was partially supported by EPSRC Programme Grant EP/I031022/1 (Natural Speech Technology).

7. REFERENCES

[1] Simon King and Vasilis Karaiskos, "The Blizzard Challenge 2010," in *Proc. Blizzard Challenge*, Kansai Science City, Japan, 2010.

[2] Simon King and Vasilis Karaiskos, "The Blizzard Challenge 2011," in *Proc. Blizzard Challenge*, Turin, Italy, 2011.

[3] Simon King and Vasilis Karaiskos, "The Blizzard Challenge 2012," in *Proc. Blizzard Challenge*, Portland, USA, 2012.

[4] Simon King, "Measuring a decade of progress in text-to-speech," *Loquens*, vol. 1, no. 1, 2014.

[5] Jianhua Tao, Keikichi Hirose, Keiichi Tokuda, Alan W. Black, and Simon King, "Introduction to the issue on statistical parametric speech synthesis," *IEEE Journal of Selected Topics in Signal Processing*, vol. 8, no. 2, pp. 170–172, 2014.

[6] H. Zen and T. Toda, "An overview of Nitech HMM-based speech synthesis system for Blizzard challenge 2005," in *Proc. of Interspeech*, 2005, pp. 93–96.

[7] K. Tokuda, Y. Nankaku, T. Toda, H. Zen, J. Yamagishi, and K. Oura, "Speech synthesis based on hidden Markov models," *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1234–1252, May 2013.

[8] Toda Tomoki and Keiichi Tokuda, "A speech parameter generation algorithm considering global variance for HMM-based speech synthesis," *IEICE TRANSACTIONS on Information and Systems*, vol. 90, no. 5, pp. 816–824, 2007.

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

[10] Keiichi Tokuda, Takayoshi Yoshimura, Takashi Masuko, Takao Kobayashi, and Tadashi Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," in *ICASSP 2000*. IEEE, 2000, vol. 3, pp. 1315–1318.

[11] Korin Richmond, "Trajectory mixture density networks with multiple mixtures for acoustic-articulatory inversion," in *Advances in Nonlinear Speech Processing*, pp. 263–272. Springer, 2007.

[12] Heiga Zen, Keiichi Tokuda, and Alan W. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039 – 1064, 2009.

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

[14] Thomas Merritt, Tuomo Raitio, and Simon King, "Investigating source and filter contributions, and their interaction, to statistical parametric speech synthesis," in *Proc. Interspeech*, 2014.

[15] Gustav Eje Henter, Thomas Merritt, Matt Shannon, Catherine Mayo, and Simon King, "Measuring the perceptual effects of modelling assumptions in speech synthesis using stimuli constructed from repeated natural speech," in *Proc. Interspeech*, 2014.

[16] V. Wan, J. Latorre, K. Yanagisawa, N. Braunschweilers, L. Chen, M. Gales, and M. Akamine, "Building HMM-TTS models on diverse data," *IEEE Journal of Selected Topics in Signal Processing*, vol. 8, no. 2, 2014.

[17] K. Tokuda, T. Kobayashi, T. Masuko, and S. Imai, "Mel-generalized cepstral analysis - a unified approach to speech spectral estimation," in *ICSLP*, 1994.

[18] S Imai, T Kobayashi, K Tokuda, T Masuko, K Koishida, S Sako, and H Zen, "Speech signal processing toolkit (SPTK), version 3.3," 2009.

[19] P. J B Jackson and C.H. Shadle, "Pitch-scaled estimation of simultaneous voiced and turbulence-noise components in speech," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 7, pp. 713–726, Oct 2001.

[20] Hanna Silén, Elina Helander, Jani Nurminen, and Moncef Gabbouj, "Ways to implement global variance in statistical speech synthesis," in *Proc. Interspeech*, 2012.

[21] H. Zen and M.J.F Gales, "Decision tree-based context clustering based on cross validation and hierarchical priors," in *Proc. ICASSP*, 2011, pp. 4560–4563.

[22] J. Latorre, M.J.F. Gales, S. Buchholz, K. Knill, M. Tamura, Y. Ohtani, and M. Akamine, "Continuous F0 in the source-excitation generation for HMM-based TTS: Do we need voiced/unvoiced classification?," in *Proc. ICASSP*, 2011, pp. 4724–4727.

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

[24] Ingwer Borg and Patrick J.F. Groenen, *Modern Multidimensional Scaling*, Springer, 2005.

[25] Joseph B Kruskal and Myron Wish, *Multidimensional scaling*, vol. 11, Sage, 1978.