8th ISCA Speech Synthesis Workshop • August 31 – September 2, 2013 • Barcelona, Spain

Investigating the shortcomings of HMM synthesis

Thomas Merritt, Simon King

The Centre for Speech Technology Research, The University of Edinburgh, U.K.
T.Merritt@ed.ac.uk, Simon.King@ed.ac.uk

Abstract

This paper presents the beginnings of a framework for formal testing of the causes of the current limited quality of HMM (Hidden Markov Model) speech synthesis. This framework separates each of the effects of modelling to observe their independent effects on vocoded speech parameters in order to address the issues that are restricting the progression to highly intelligible and natural-sounding speech synthesis.

The simulated HMM synthesis conditions are performed on spectral speech parameters and tested via a pairwise listening test, asking listeners to perform a "same or different" judgement on the quality of the synthesised speech produced between these conditions. These responses are then processed using multidimensional scaling to identify the qualities in modelled speech that listeners are attending to and thus forms the basis of why they are distinguishable from natural speech.

The future improvements to be made to the framework will finally be discussed which include the extension to more of the parameters modelled during speech synthesis.

Index Terms: Speech synthesis, Hidden Markov models, Vocoding

1. Introduction

Despite several years of improvements in the quality of speech generated using HMM (Hidden Markov Model) synthesis, this type of synthetic speech still stubbornly remains significantly less natural than speech output from good concatenative (unit selection) synthesis systems [1, 2], as consistently reflected in the results from the annual Blizzard challenge [3, 4, 5]. Although it can achieve higher intelligibility than unit selection, HMM synthesis is not yet as natural as unit selection, and neither are judged by listeners to be as natural as real speech.

It is common in the literature to find the cause for the reduced naturalness of HMM speech stated as "over-smoothing", and that this is the fault of the statistical model, but to the best of our knowledge there are no formal, published studies supporting this claim. The idea of "over-smoothing" is at first glance seemingly a simple one, but may conflate a number of different effects of signal representation and of statistical modelling in both spectral and temporal domains. Smoothing is inherent in the statistical modelling framework, of course. The spectral envelope is smoothed first by the low-dimensional representation, then again by averaging over consecutive frames and over multiple tokens. The temporal structure of the speech parameters is smoothed because the model represents the trajectory with limited resolution (e.g., 5 states per phone-sized-unit).

What is needed is a framework in which we can separate out the different contributions of the various processes of modelling. This is the contribution of this paper.

1.1. A simulation framework

This paper introduces such a framework and – as a first illustration of its use – tests a couple of the potential causes of the degradation in naturalness introduced by the use of statistical models. The framework is general and could be applied to many different aspects of the problem. The idea is to *simulate* the effects of modelling vocoded speech, in a carefully controlled manner. Knowledge obtained by such experiments could then be used to identify those areas that are causing the problem, and to eventually rectify them.

Current HMM-based synthesisers are large, complex systems. There are interactions between the signal processing (e.g., how the spectral envelope is extracted and how it is represented for the purposes of modelling) and the modelling (e.g., the parameter sharing structure of the model and how much data are available to estimate each free parameter) which need to be investigated. In the work presented here, this will be done by removing the modelling part completely and replacing it with a series of operations which are designed to simulate some modelling effects. Our proposed approach allows us to vary the strength of these effects, and to examine the interactions between them. Thus, by using simulation, we can continuously vary the system from being a simple vocoder at one end of the scale, to a simulated HMM synthesiser at the other. In this paper, the effects that we use are temporal smoothing and variance scaling of the speech parameters representing the spectral envelope.

1.2. Measuring the effects

The second component of the proposed framework is perceptual testing of the acoustic consequences of the simulated effects of statistical modelling. Asking listeners to attend to specific aspects of the speech is problematic [6, 7] and also risks biasing them towards certain phenomena. Since we are not entirely sure what perceptual dimensions listeners use when rating the naturalness of synthetic speech, it is not clear what aspects of the signal we could ask them to attend to. Therefore, we adopt a less direct methodology, and ask the listeners to perform a very simple task where the instructions contain no bias towards any particular acoustic property or perceptual dimension. This task is a simple "same or different" judgement on pairs of stimuli, from which we can derive a matrix of pairwise perceptual distances. Multidimensional scaling (MDS) allows such data to be visualised and from this visualisation we can identify the perceptual dimensions, that is, what the listeners are attending to. Tracing these back to the simulated effects involves interpreting the MDS visualisation.

1.3. Structure of this paper

Section 2 will discuss how we implemented a simulation of HMM synthesis, section 3 will introduce the method for perceptually testing the speech created under this simulation, then section 4 presents the results from this testing. Based on these results, we offer an interpretation and some conclusions in section 5 followed by a summary of the contributions of this paper. Finally, section 6 will suggest future work, including how we plan to use the proposed framework to simulate many more of the effects of statistical modelling.

2. Methodology

Our aim is to tease apart the complex effects of statistical modelling on synthetic speech. In order for the contributing factors (to shortcomings in the quality of speech output by HMM synthesis) to be investigated, we need a framework in which these effects can be individually manipulated – a kind of 'oracle' HMM synthesiser which allows for complete control over each aspect of the system, varying it between some form of 'ideal', or 'perfect' component and the real component used in a full HMM synthesiser. An obvious example of the 'ideal' is a vocoder, which has access to natural speech parameters and is so unaffected by any flaws in the way the statistical modelling part reconstructs these.

2.1. Scope of the current investigation

In the present work, we concentrate on global simulations of the statistical modelling part of the system. This is illustrated in figure 1, where we can see that the speech parameter extraction and waveform generation (reconstruction) parts are the same as in a full HMM synthesiser. Extraction of the spectral, F0, and aperiodic energy speech parameters is performed as usual, with the use of STRAIGHT (Matlab implementation)[1] [8, 9] followed by SPTK [10] to convert the spectral envelope to line spectral frequencies (LSFs), F0 to log F0 and aperiodic energy to band aperiodic energy. We chose to use LSFs because they are more convenient for visualisation than, say, Mel-generalised cepstra, and this should ease the interpretation of the results later. The conversion of F0 to log F0 and aperiodic to band aperiodic was also performed to simulate common modelling conditions of all speech parameters, this allows us to better track the effect that modelling has on the spectral envelope parameters by implementing a system which is more realistic. We also focus only on the spectral envelope speech parameters here; experimentation with the other speech parameters is future work.

Following the application of our modelling simulations, the LSFs, log F0 and band aperiodic energy parameters were converted back into spectral, F0 and aperiodic energy speech parameters using SPTK [10] before performing the 'reconstruction' phase of HMM speech synthesis, by inputting the speech parameters into STRAIGHT (Matlab implementation) to obtain the synthesised speech waveform as output.

2.2. Simulating "over-smoothing"

There are several ways in which the output speech parameters of an HMM synthesiser are "too smooth". Here, we concentrate on temporal effects, leaving spectral smoothness as future work. Looking at the output of typical HMM systems [2, 11], we generally find far less temporal detail than is observed in the speech parameters for natural speech. Some of this detail may simply be noise introduced by the spectral envelope estimation process, but some of it may be perceptually important. We investigate this by temporally smoothing the speech parameters, which simulates the limited temporal resolution of 5-state-per-phone models and the subsequent MLPG [1, 12] trajectory generation algorithm.

[1] STRAIGHT V40_007 methods were used, these were written by Hideki Kawahara

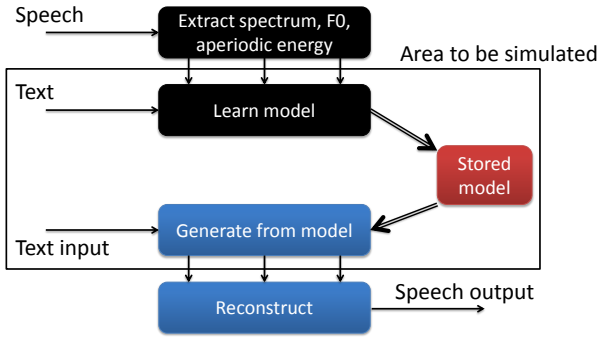

Figure 1: *Training and using an HMM speech synthesiser, illustrating the part of the process that is simulated here.*

Another consequence of statistical modelling is that the variance of the generated speech parameters is lower than those from natural speech. This has long been known to significantly reduce the quality of the generated speech and is why mitigating this by considering Global Variance (GV) [13, 2] has such as dramatic positive effect on quality. However, GV cannot guarantee to perfectly restore the correct variance of the parameters. We simulate the effect of modelling and of GV by scaling the standard deviation of the speech parameters by a value greater or less than 1.0.

Removing temporal detail via smoothing will also slightly reduce the variance of the speech parameters. We can examine the interaction between temporal smoothness and variance by applying both effects, with varying strengths. It is worth repeating at this point that temporal smoothing and variance scaling are certainly not a comprehensive simulation of HMMs synthesis, but that they were used here as a starting point for an ongoing investigation and that more complex effects will be investigated in future work.

The effects simulated in the current work are all applied to each speech parameter independently and are implemented utterance-by-utterance.

2.2.1. Temporal smoothing

The smoothing effect was implemented as a weighted moving average, sliding a Hanning window over the signal (i.e., each LSF in turn), to simulate the limited temporal resolution of HMM modelling. The width of the window was varied, to impose varying amounts of smoothing.

2.2.2. Variance scaling

Variance adjustment was implemented as a simple scaling of the standard deviation by a fixed factor. For each parameter (i.e., each LSF) in turn, the mean value over the utterance

was found and subtracted before multiplying the parameter by a scalar value, and finally adding the mean back in. By altering the scalar value, the standard deviation is correspondingly adjusted, to simulate both reduced variance (which is commonly observed in HMM synthesis) and increased variance (e.g., as may happen if a Gaussian p.d.f. is poorly estimated during training, or when GV fails to re-instate the appropriate amount of variance).

3. Experiments

A range of simulated effects were selected to be tested, with the strengths of modifications being selected by informal listening to reflect the sorts of imperfections we have ourselves encountered in many of the HMM synthesis systems we have built. For the temporal smoothing, Hanning window sizes of 80 and 110 frames (at a frame rate of 5 msec) were selected, along with a 'no smoothing' condition. Smaller window widths (i.e., less smoothing) were found to produce negligible perceptual effects. Variance adjustment involved scaling the standard deviation by scalar values of 0.6, 0.8, 1.2 and 1.4 as well as a 'no variance adjustment' condition equivalent to scaling by 1.0. These particular values for smoothing and variance adjustment were selected to provide audibly different speech quality, whilst staying within the range of qualities that we have observed in real HMM synthesisers.

3.1. Materials

The speech corpus used for testing was a set of Harvard Sentences [14] read by a male professional speaker of British English (known as 'Nick' and whose speech has been used in the Hurricane Challenge [15] and who also features in the 'mngu0' acoustic-articulatory corpus[2] [16]), this was sampled at 16 KHz. The methodology for preparing the stimuli was, as described above, to extract speech parameters using STRAIGHT and SPTK, to apply the two simulated effects of smoothing and variance adjustment with all possible combinations of strengths including the 'no modification' conditions, then to reconstruct the waveform. Order 30 LSF coefficients were used as this offers a good representation of the spectral information for the speech at the sampling rate used. The result was $3 \times 5 = 15$ versions of each of 40 sentences.

The variance adjustment method was applied per speech parameter per utterance independently, so the mean speech parameter value subtracted before scaling is influenced by the amount of silence present; therefore, the material was manually edited to leave only just a few 100 msec of leading and trailing silence. Care was also taken to remove any background noise present during the non-speech, because in preliminary experiments this became perceptually much more apparent after applying some of modifications.

3.2. Listening test

In the listening test, listeners had to make forced choice 'same or different quality' judgements about pairs of stimuli.

The testing was performed by applying each of the 15 simulation conditions (called A to O) as defined in table 2, which combine smoothing and/or variance adjustment to each of the 40 sentences. The 40 sentences were divided into 20 pairs (sentences 1 & 2, sentences 3 & 4, and so on), and for each of these pairs of sentences, all possible combinations of conditions (e.g., sentence 1 in condition A + sentence 2 in condition F) were created, except for pairs of identical conditions (e.g., sentence 1 in condition A + sentence 2 in condition A), as shown in figure 3.

This resulted in $20 \times ((15 \times 15) - 15) = 4200$ pairs of sentences, which were then randomised in order and divided amongst 30 listeners, resulting in each listener listening to 140 pairs of sentences and thus making 140 'same or different' judgements. These listeners were selected at random from applicants to an online advert placed in the University of Edinburgh's Student And Graduate Employment service; all were native English speakers with no self-reported hearing problems. The stimuli pairs were presented in a randomised order per listener over high quality headphones in quiet sound-proofed booths with no distractions.

[2] http://www.mngu0.org

Condition index	Hanning smoothing window size	Standard deviation scaling
A	none	0.6
B	80	0.6
C	110	0.6
D	none	0.8
E	80	0.8
F	110	0.8
G	none	none
H	80	none
I	110	none
J	none	1.2
K	80	1.2
L	110	1.2
M	none	1.4
N	80	1.4
O	110	1.4

Figure 2: *The 15 conditions combining each level of smoothing (including no smoothing) and each amount of standard deviation scaling (including no modification)*

		Sentence 1														
		A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
Sentence 2	A	×	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	B	✓	×	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	C	✓	✓	×	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	D	✓	✓	✓	×	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	E	✓	✓	✓	✓	×	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	F	✓	✓	✓	✓	✓	×	✓	✓	✓	✓	✓	✓	✓	✓	✓
	G	✓	✓	✓	✓	✓	✓	×	✓	✓	✓	✓	✓	✓	✓	✓
	H	✓	✓	✓	✓	✓	✓	✓	×	✓	✓	✓	✓	✓	✓	✓
	I	✓	✓	✓	✓	✓	✓	✓	✓	×	✓	✓	✓	✓	✓	✓
	J	✓	✓	✓	✓	✓	✓	✓	✓	✓	×	✓	✓	✓	✓	✓
	K	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	×	✓	✓	✓	✓
	L	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	×	✓	✓	✓
	M	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	×	✓	✓
	N	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	×	✓
	O	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	×

Figure 3: *One set of pairings of sentences and conditions in the listening test.*

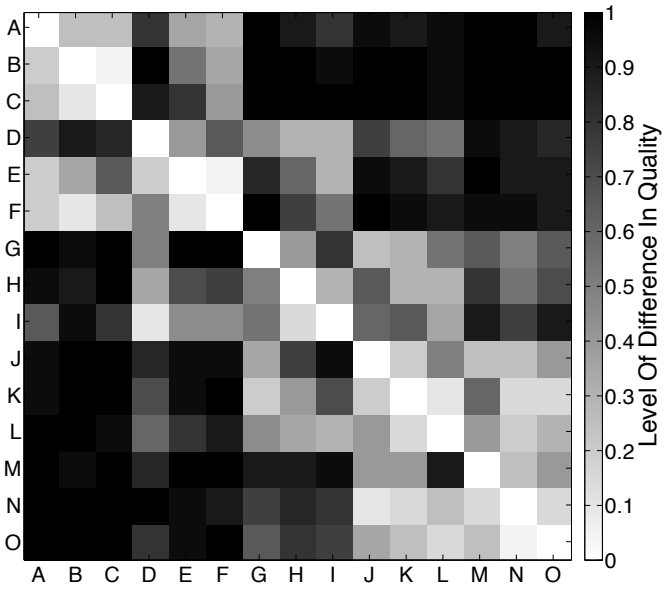

Figure 4: *Listeners' responses between conditions presented in figure 2, pooled across all sentences and listeners. Darker shades indicate greater perceived dissimilarity between conditions.*

3.3. Multidimensional scaling

The raw listener responses were pooled across all listeners and all sentences for each individual combination of modifications. The result is a dissimilarity matrix, in which each cell contains a number indicating the perceived dissimilarity between two conditions. Figure 4 shows this matrix graphically: each cell contains the number of comparisons between a pair of conditions marked as 'different' by listeners. Multidimensional scaling was used to analyse this matrix, and create a plot in which each condition appears as a point. Short distances between points on the plot indicate perceptual similarity and large distances indicate dissimilarity [17]. We used a Matlab implementation of MDS based on Kruskal's normalised STRESS1 criterion[3].

4. Results

MDS projects the dissimilarity matrix into a multi-dimensional space. In order to find an appropriate dimensionality of this space, one must compromise between accuracy of representation (in higher dimensions, the correspondence between dissimilarity and distance in the space will be more precise) against the need for a modest number of dimensions to allow for the data to be visualised and for the axes to be interpreted. The so-called stress value computed as part of the multidimensional scaling algorithm reflects this tradeoff; figure 5 plots the stress value for various dimensionalities. It seems that three dimensions is a reasonable operating point for our data.

The first two dimensions of the three-dimensional space found by multidimensional scaling is given in figure 6. Distance in this space indicates perceived dissimilarity: the closer a point is to the natural unmodified speech, the "more natural" it sounds. It is immediately apparent that the listeners judgements cannot be explained by a single dimension and that they

[3]function 'mdscale' from the Matlab statistics toolbox

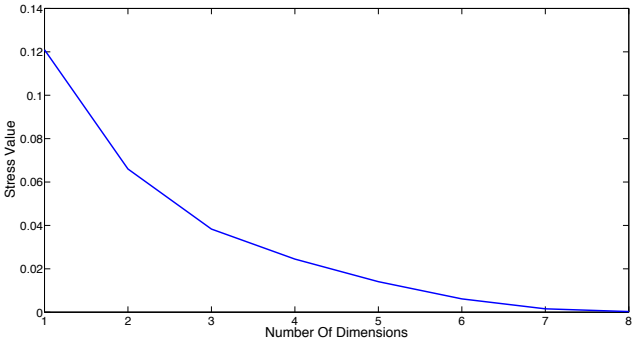

Figure 5: *Stress levels returned by MDS at different dimensions.*

are making their decisions based on more than one aspect of the speech:

- The horizontal axis seems to relate to the amount of LSF variance, with the reduced variance speech clearly different from the increased variance speech
- The vertical axis seems to relate to overall quality of synthesis, regardless of the LSF variance, with both reduced and increase variance speech being placed towards the top of the space, whereas natural speech is at the bottom.

This plot also shows that the smoothing has only a secondary effect, probably simply because it has the side effect of slightly reducing variance. When the variance is too high (right hand side of figure 6), then the smoothing has a beneficial effect, moving the points lower and therefore closer to natural speech.

5. Conclusions

We have introduced a simple-to-use, extensible methodology that can tease apart the contributions to speech quality of the various components of an HMM-based text-to-speech system. The fundamental idea is to simulate all or part of the system, and thus to gain explicit control over the system's behaviour. In this paper, we have demonstrated the use of this framework in a straightforward way, by simulating a complete HMM-based synthesiser as simply a combination of smoothed parameter trajectories and incorrect variance.

Even from this very simple simulation, we can conclude that listeners are able to perceive different types of quality reduction: the MDS analysis reveals that they can make overall quality judgements (vertical axis of figure 6) and at the same time clearly distinguish whether this is due to too high or too low variance. It also seems fairly safe to conclude that *temporal* smoothness in LSF trajectories is not really a problem and leads to only very small perceptual effects.

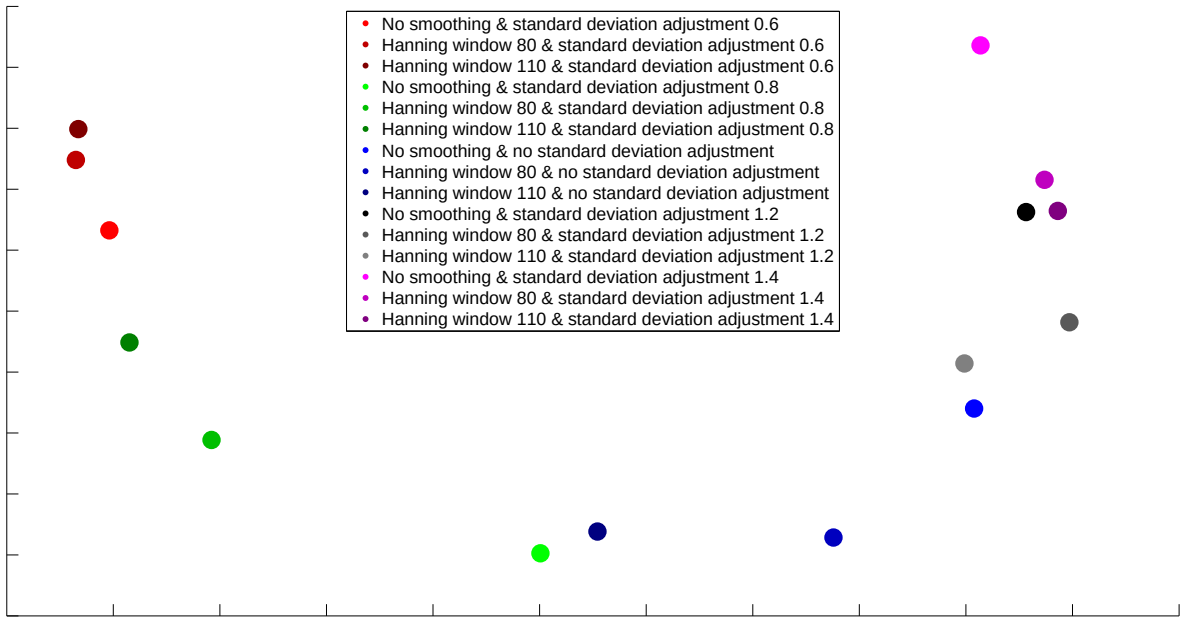

Figure 6: *Plot of the first two dimensions of the three-dimensional space found using MDS.*

6. Future work

The next steps are obvious: to extend the range of simulated effects of modelling and conduct further listening tests followed by MDS analysis of the responses. The ultimate aim is a system that can be continuously controlled between an 'oracle' vocoder and a fully-modelled text-to-speech system. Some categories of effects that we would like to simulate next include:

- *spectral envelope* over-smoothness: formant dulling and sharpening; suppression or emphasis of spectral detail
- averaging across *multiple tokens* of similar speech sounds (e.g., phonemes in context) at frame, state and model granularities
- poor modelling of the *covariance* within a set of speech parameters (e.g., LSFs), resulting in inconsistent sets of values
- *inconsistencies* between the different speech parameter streams (e.g., aperiodic energy vs. spectral envelope) caused by use of different model parameter tying structures
- model boundary *discontinuities* in the trajectory (which may be disguised but not overcome by MLPG) occurring at transitions between HMMs of phoneme-sized units

7. Acknowledgements

This work was supported by EPSRC Programme Grant EP/I031022/1 (Natural Speech Technology). Thanks also to Heng Lu for his assistance with STRAIGHT and SPTK, Rob Clark for his advice on MDS and Cassie Mayo for her advice on perceptual testing and experimental design.

8. References

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

[2] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039 – 1064, 2009. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167639309000648

[3] S. King and V. Karaiskos, "The Blizzard Challenge 2011," in *Proc. Blizzard Challenge*, Turin, Italy, 2011.

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

[5] ——, "The Blizzard Challenge 2009," in *Proc. Blizzard Challenge*, Edinburgh, United Kingdom, 2009.

[6] 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.

[7] ——, "Multidimensional scaling of listener responses to synthetic speech," 2005.

[8] 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 communication*, vol. 27, no. 3, pp. 187–207, 1999.

[9] C. Liu and D. Kewley-Port, "STRAIGHT: A new speech synthesizer for vowel formant discrimination," *Acoustics Research Letters Online*, vol. 5, p. 31, 2004.

[10] S. Imai, T. Kobayashi, K. Tokuda, T. Masuko, K. Koishida, S. Sako, and H. Zen, "Speech signal processing toolkit (SPTK), version 3.6," 2012.

[11] 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, 2013.

[12] T. Yoshimura, "Simultaneous modeling of phonetic and prosodic parameters, and characteristic conversion for HMM-based Text-To-Speech systems," Ph.D. dissertation, Ph. D. thesis, Nagoya Institute of Technology, 2002.

[13] T. Tomoki and K. 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.

[14] IEEE, "IEEE recommended practice for speech quality measurement," vol. 17, no. 3, pp. 225 – 246, sep 1969.

[15] M. Cooke, C. Mayo, and C. Valentini-Botinhao, "Intelligibility-enhancing speech modifications: the Hurricane Challenge," in *Proc. Interspeech*, Lyon, France, 2013.

[16] K. Richmond, P. Hoole, and S. King, "Announcing the electromagnetic articulography (day 1) subset of the mngu0 articulatory corpus," in *Proc. Interspeech*, Florence, Italy, August 2011, pp. 1505–1508.

[17] I. Borg and P. J. Groenen, *Modern Multidimensional Scaling*. Springer, 2005.