Intelligibility analysis of fast synthesized speech

Cassia Valentini-Botinhao[1], *Markus Toman*[2], *Michael Pucher*[2], *Dietmar Schabus*[2], *Junichi Yamagishi*[1,3]

[1] The Centre for Speech Technology Research (CSTR), University of Edinburgh, UK
[2] Telecommunications Research Center Vienna (FTW), Austria
[3] National Institute of Informatics, Japan

{cvbotinh,jyamagis}@inf.ed.ac.uk, {toman,pucher,schabus}@ftw.at

Abstract

In this paper we analyse the effect of speech corpus and compression method on the intelligibility of synthesized speech at fast rates. We recorded English and German language voice talents at a normal and a fast speaking rate and trained an HSMM-based synthesis system based on the normal and the fast data of each speaker. We compared three compression methods: scaling the variance of the state duration model, interpolating the duration models of the fast and the normal voices, and applying a linear compression method to generated speech. Word recognition results for the English voices show that generating speech at normal speaking rate and then applying linear compression resulted in the most intelligible speech at all tested rates. A similar result was found when evaluating the intelligibility of the natural speech corpus. For the German voices, interpolation was found to be better at moderate speaking rates but the linear method was again more successful at very high rates, for both blind and sighted participants. These results indicate that using fast speech data does not necessarily create more intelligible voices and that linear compression can more reliably provide higher intelligibility, particularly at higher rates.

Index Terms: fast speech, HMM-based speech synthesis, blind users

1. Introduction

Blind individuals are capable of understanding speech reproduced at considerably high speaking rates [1]. As screen readers become an essential computer interface for blind users, a challenge arises: how to provide intelligible synthesized speech at such high rates? The standard HSMM-based synthesizer [2] models speech duration by using explicit state duration distributions but for very fast speaking rates this is often not sufficient [3]. It is also unclear whether using fast speech to train a synthesizer can create more intelligible fast synthesized speech than other sorts of compression methods.

Fast speech production and perception has been the target of various studies [4–8]. When producing fast speech vowels are compressed more than consonants [4] and both word-level [5] and sentence-level [6] stressed syllables are compressed less than unstressed ones. Yet another important aspect of fast speech is the significant reduction of pauses. It is claimed that reducing pauses is in fact the strongest acoustic change when speaking faster [7], most probably due to the limitations of how much speakers can speed up their articulation rate [8]. It is argued that these observed changes are the result of an attempt to preserve the aspects of speech that carry more information. The presence of pauses however have been shown to contribute to intelligibility [9].

It has been shown that fast speech (around 1.56 times faster than normal speech) is harder to process, in terms of reaction time, and also preferred less than linearly compressed speech [5, 10]. Linearly compressed speech was found to be more intelligible and better liked than a nonlinearly compressed version of speech where fast speech prosodic patterns were mimicked [5]. The author claims that possibly the only nonlinear aspect of natural fast speech duration changes that can improve intelligibility at high speaking rates is pause removal but only when rates are relatively high [10]. Another nonlinear compression method is the MACH1 algorithm [11]. This method is also based on the acoustics of fast speech with the addition of compressed pauses. It has been shown that at high speaking rates (2.5 and 4.1) MACH1 improves comprehension and is preferable to linearly compressed speech but no advantage was found at the fast speech speaking rate (1.4) [12].

Fast synthesized speech generated by a formant-based system was found to be less intelligible than fast natural speech and the intelligibility gap grows with the speaking rate [13]. More recently the authors in [14] evaluated the intelligibility of a wider range of synthesizers: formant, diphone, unit selection and HMM-based. It was found that the unit selection systems were more intelligible across speech rates. In this evaluation, however, the evaluated synthesizers were based on different speakers and the compression methods adopted by each system were not reported. Literature on fast synthesized speech also focuses on the effect on blind listeners. To improve duration control of HMM-based systems for blind individuals [3] proposed a model interpolation method. Pucher et al. found that interpolating between a model trained with normal and a model trained with fast speech data results in speech that is more intelligible and preferable, for both blind and non blind individuals.

In this paper, we are interested in analysing two aspects of fast synthesized speech. First, the corpus used to train synthesis models, i.e., is it really necessary or even helpful to use fast speech recordings? Second, compression method; which is more effective: a nonlinear manipulation of speech duration or a linear compression method? We evaluate intelligibility of a fast and a normal female Scottish voice and a German male voice, compressed using two nonlinear and one linear method and presented to listeners at different rates.

This paper is organized as follows: Section 2 describes the methods used to create synthetic speech at fast rates, Section 3 presents the corpus used for training the synthesis models and details on how models were trained, Section 4 shows the design and results of intelligibility listening experiments, Section 5 presents a discussion on these results followed by conclusions in Section 6.

2. Compression methods

In this section, we describe methods that can create synthetic speech at fast rates, referred to here as compression methods. The first two methods we describe manipulate the state duration model parameters (mean and/or variance) while the third is applied to the synthesized speech waveform. The first two methods are considered to be nonlinear as each state is compressed at a different rate, as opposed to the third method, which is a linear method that compresses the waveform uniformly across time.

2.1. Variance scaling

Variance scaling is the standard method for duration control in HSMM-based synthesis [15]. With this method we compute the duration of state i as:

$$d_i = \mu_i + \rho\sigma_i \quad (1)$$

where μ_i and σ_i are the mean and variance of the state duration model and ρ is a factor that controls the variance scaling. When $\rho = 0$ the duration is set to the mean state duration, $\rho > 0$ makes synthetic speech slower and $\rho < 0$ faster. The scaling factor is fixed across all states. State duration control is then proportional only to the variance: states whose duration model variance is higher will be compressed more. With this method we can potentially capture certain non-linearities between normal and fast speech durations.

2.2. Model interpolation and extrapolation

In previous work with fast synthetic speech [3], we showed that model interpolation [16, 17] can outperform the variance scaling method in terms of intelligibility and listener preference. Given two voice models of the same speaker trained with speech recorded at normal and fast speaking rates, the most successful method in that study was one that applied interpolation between duration models, using the normal speaking rate models of cepstral, fundamental frequency and aperiodicity features. The interpolated duration d_i for state i is calculated as:

$$d_i = (1-\alpha)\,\mu_i^n + \alpha\,\mu_i^f \quad (2)$$

where μ_i^n and μ_i^f denote the mean duration of state i in the normal and fast duration model and α is the interpolation ratio to control the speaking rate. We can generate speaking rates beyond the rate of the fast model by extrapolating ($\alpha > 1$).

For the experiments in the present paper, we have implemented an additional constraint in this method. It is possible that for a given state of a given phone, the mean duration μ_i^f from the fast model is actually longer than the mean duration μ_i^n of the normal model, causing the speech segments generated for this state to become *slower* with growing α. If this is the case, we do not interpolate or extrapolate, but apply a linear factor β to μ_i^n, where β reflects the overall mean speaking rate difference between the normal and the fast voice models ($\beta = 1/1.55$ in our experiments).

2.3. WSOLA

The waveform similarity overlap and add (WSOLA) method proposed in [18] was chosen here to illustrate the effect of a linear compression. The method provides high enough quality while being computationally efficient and robust [18]. In WSOLA speech frames to be overlapped are first cross-correlated to provide an appropriate time shift that ensures frames are added coherently, inspired by the idea that modified speech should maintain maximum local similarity to the original signal.

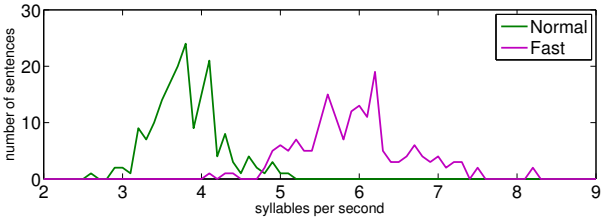

Figure 1: *English TTS voices: syllables per second distribution.*

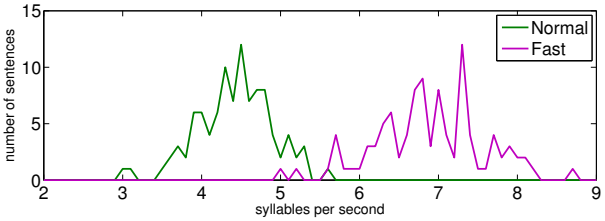

Figure 2: *German TTS voices: syllables per second distribution.*

3. Speech databases and voices

We present the English and German corpora used in our experiments as well as details of how we trained the synthetic voices.

3.1. English – corpus and voices

We recorded a Scottish female voice talent reading 4600 sentences at a normal speed and 800 sentences at a fast speed with the instruction to speak as fast as possible while maintaining intelligibility.

To train the acoustic models, we extracted the following features from the natural speech sampled at 48 kHz: 59 Mel cepstral coefficients [19], Mel scale fundamental frequency F0 and 25 aperiodicity band energies extracted using STRAIGHT [20]. We used a hidden semi-Markov model as the acoustic model. The observation vectors for the spectral and excitation parameters contained static, delta and delta-delta values. One stream was set for the spectrum, three for F0 and one for aperiodicity.

We trained two voices. What we refer to as the model N, is a voice trained only with speech produced at the normal speaking rate. This model was adapted [21] using the 800 sentences of fast speech to create what is referred to as the voice F.

To measure the speaking rate of each synthetic voice we calculated the rate of syllables per second (SPS) and words per minute (WPM) for each sentence used in the evaluation. On average the SPS values of the normal and the fast voice are 3.8 and 6.0 while the values for WPM are 206.7 and 320.9, respectively. Speech synthesized using the fast model is around 1.55 times faster, which agrees with the literature [5] on naturally produced fast speech. Fig. 1 shows the histogram of SPS across synthesized sentence for each voice.

3.2. German – corpus and voices

We used a very similar setup to record and train the German voice. We recorded an Austrian German voice talent reading

W-N	WSOLA applied to normal speech
W-F	WSOLA applied to fast speech
V-N	Variance scaling applied to model N
V-F	Variance scaling applied to model F
I	Interpolation of model N and F

Table 1: *Methods evaluated.*

4387 sentences at a normal and 198 sentences at a fast speaking rate. The German recordings were sampled at 44.1 kHz and we extracted 39 Mel cepstral coefficients. The F voice was obtained by adapting the duration model only as fewer sentences were available for adaptation [3]. Otherwise the procedure and parameters were the same as for English.

The average SPS values for the normal and fast German synthetic voices are 4.5 and 7.0, and the WPM values are 152.7 and 237.1. The German voice is thus considerably faster than the English voice, at both speaking rates. Interestingly, the fast model is also about 1.55 times faster than the normal model, i.e., the speed-up factor between the two English models and between the two German models is the same. Fig. 2 shows the SPS distribution for the two German models.

4. Evaluation

We conducted two listening experiments with the English voices, one using natural speech and the other TTS; while for the German data only the TTS voices were evaluated, but by both blind and sighted individuals.

We evaluate intelligibility at four different speaking rates: 1.25, fast (the speed of fast speech), 2.0 and 3.0, where numbers refer to speed increase with respect to the normal voice calculated sentence by sentence, remembering here that fast speech is around 1.55 times faster than normal speech. Rates were chosen to reflect conversational, fast and two ultra fast speeds.

The methods we evaluate are presented in Table 1 [1]. Not all methods are evaluated at all speaking rates, for instance at rates smaller or equal to the fast rate W-F, V-F and I were not evaluated. To generate compressed samples using the variance and the interpolation methods it was necessary to progressively change the scale factor to obtain the desired duration. The implementation of WSOLA used here was provided as support material for [22].

Results are presented as percentage of word errors, calculated per listener as the percentage of words that were not transcribed, misspellings taken into account.

4.1. English – evaluation

We evaluate the intelligibility of natural speech compressed only with the WSOLA algorithm as the other two methods can not be applied directly to natural speech. We compare two natural speech compressions: W-N and W-F, compression applied to the normal and the fast speech databases.

For the TTS evaluation, we compare the three different compression methods described in Section 2, although not all methods were evaluated for all speaking rates.

4.1.1. Listening experiment

We performed two listening experiments, one with natural speech and the other with the TTS voices. Each experiment was performed by 20 native English speakers without TTS expertise. Each participant transcribed 10 different sentences for each of the tested methods. The natural speech sentences were selected from news articles while for the TTS experiments sentences were chosen from the first few sets of the Harvard dataset [23].

4.1.2. Results

Fig. 3 shows the percentage of word errors for each speaking rate obtained in the natural (blue) and TTS (red) experiments.

We can see that the TTS voices created using WSOLA are the most intelligible across all tested speaking rates and that this advantage grows with increasing speaking rate. At the fastest rate the TTS voice W-N results in less than 20 % word errors while the word errors obtained by V-N, V-F and I are higher than 40 %, i.e., errors doubled. Interpolation is slightly better than variance scaling, although not significantly. Although not reported here, we have also observed the WSOLA advantage when F is obtained by adapting the duration model only.

Word errors are smaller when compressing speech synthesized from the normal model (W-N) as opposed to a fast model (W-F), as results for speaking rate 2xs show. Although differences are not significant, error levels for by V-F and I are slightly smaller than V-N at all speaking rates. At the fast speaking rate, we can see that the fast voice is less intelligible than the normal voice with linear compression applied.

Compared to the natural speech results (in blue) we can see that error scores are significantly higher for TTS voices. The increase in error seen for W-F compared to W-N for TTS voices can also be observed for natural speech, pointing to the fact that the fast natural speech is also less intelligible than linearly compressed normal speech.

4.2. German – evaluation

A similar evaluation was carried out for German to assess the intelligibility achieved by the methods described in Section 2.

4.2.1. Listening experiment

For the German data, only TTS voices were evaluated. The participants in the listening test consisted of two groups: 16 blind or visually impaired participants, 15 of whom reported using TTS in their everyday life, and 16 sighted participants with no TTS expertise. Each participant transcribed 100 different sentences such that within a participant group, every combination of method and speaking rate was evaluated once. The sentences were selected from news articles and parliamentary speeches.

4.2.2. Results

The results are shown in Fig. 4, where the two bars per condition reflect the results from the two participant groups. As expected, the blind listeners (yellow) generally achieve lower word error percentages than the sighted listeners (red).

Similar to the English results, WSOLA compression of speech synthesized from the normal model (W-N) is the best method overall. However, up to 2xs, both WSOLA of fast speech (W-F) and interpolation (I) yield results competitive to W-N. At the "fast" rate, where W-F, I and also V-F, are equivalent to simply the fast voice model, these methods even achieve significantly better results than W-N for the sighted listeners. At the "fast" and 2xs rates, W-F and I perform significantly better than variance scaling of the normal model (V-N), confirming the results of [3]. However, we see a very clear advantage of the WSOLA methods at the fastest rate 3xs, where the error

[1] Speech samples used in the evaluation can be found at: http://wiki.inf.ed.ac.uk/CSTR/SalbProject

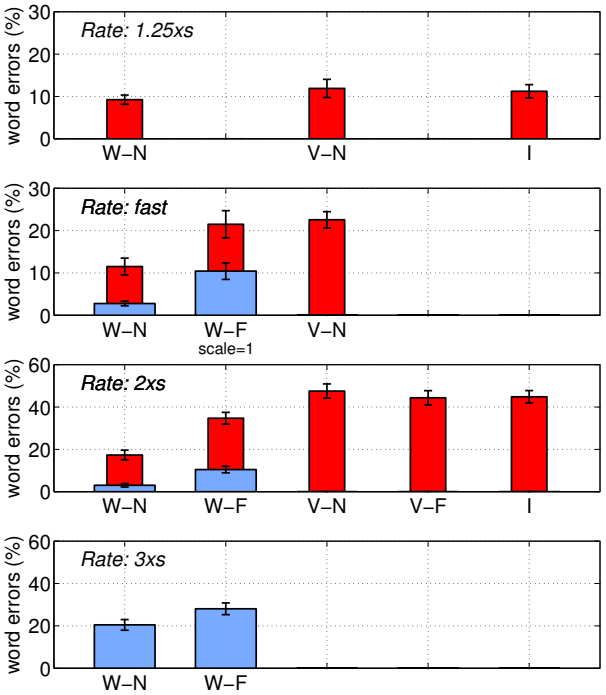

Figure 3: *English results: TTS (red) and natural speech (blue).*

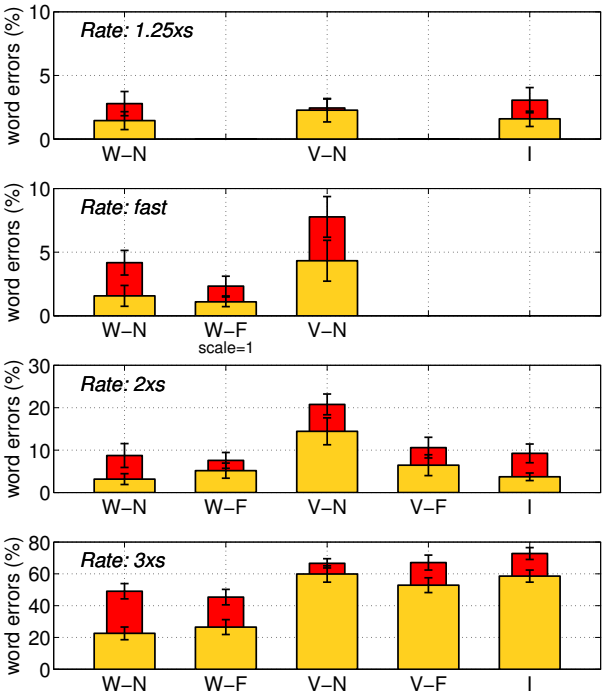

Figure 4: *German results: Sighted (red) and blind/visually impaired listeners (yellow).*

percentages of V-N, V-F and I are much higher, yielding a picture similar to the English results at 2xs. There is no significant difference between W-N and W-F at the 2xs and 3xs rates.

5. Discussion

As found in other studies of natural fast speech [5, 10], our results using the English data also indicate that linear compression can produce more intelligible voices than nonlinear methods based on or directly derived from the acoustics of fast speech. English results show that there is no additional advantage to using recordings of fast speech to build a synthetic voice and it is possible to maintain intelligibility at higher speaking rates by applying a simple linear compression method to the synthesized waveform. This is supported by results with the natural speech corpus, where we also found that fast natural speech is not as intelligible as linearly compressed normal speech.

The German results tell a slightly different story. There we also see that linear compression is beneficial at very high speaking rates (3xs) compared to interpolation and variance scaling. For lower rates (2xs), we find that interpolation is equally good as linear compression. This indicates a potential use of a combined method of interpolation for fast speaking rates and linear compression for ultra-fast speaking rates. We hypothesize that different results were found for the German data due to the inherent higher intelligibility of the German fast speech, which can also be seen in the performance differences of linear compression of synthesized speech from fast models (W-F) which performs better for the German data. We want to investigate this hypothesis in the future by carrying out a detailed analysis of fast speech durations from different speakers. Concerning the performance of blind listeners we can confirm results presented in previous studies [1, 3], which show that blind listeners achieve lower word-error-rates than non-blind listeners.

Considering results on both databases we hypothesize that methods that use recordings of fast speech such as adaptation or interpolation are perhaps only as intelligible as the fast data they use. Relying on having fast speech that is intelligible enough is challenging as this data is quite difficult to produce considering that both of our speakers are voice talents. Using more recordings of fast speech is also not helpful as more fast sentences were used for the English voices. Moreover it is not yet clear how to reach very high speaking rates with model interpolation and adaptation as these methods are limited by the fact that no skip is allowed. The weak performance of the variance scaling method for fast speaking rates (2xs, 3xs) is in agreement with the poor results obtained by HMM-based voices in [14].

6. Conclusion

We showed that linear compression outperforms the variance scaling and interpolation methods for ultra-fast (3xs) speaking rates in German and English. For fast speaking rates (2xs) linear compression outperformed other methods for English while being as good as interpolation for German. In general we see that the usage of fast speech data in interpolation (I) or linear compression (W-F) is dependent on the quality of the data.

As future work, we plan to evaluate the intelligibility of the German language corpus and the TTS voices in English with blind participants as well. Additionally, we plan to analyse the acoustic properties of both fast speech corpus in more detail in order to explain the differences in their intelligibility.

7. Acknowledgement

This work was supported by the BMWF - Sparkling Science project *Sprachsynthese von Auditiven Lehrbüchern für Blinde SchülerInnen* (SALB). The Competence Center FTW is funded within the program COMET - Competence Centers for Excellent Technologies by BMVIT, BMWA, and the City of Vienna.

8. References

[1] A. Moos and J. Trouvain, "Comprehension of ultra-fast speech – blind vs. 'normally hearing' persons," in *Proc. Int. Congress of Phonetic Sciences*, vol. 1, 2007, pp. 677–680.

[2] H. Zen, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "A hidden semi-markov model-based speech synthesis system," *IEICE Trans. Inf. Syst.*, vol. E90-D, no. 5, pp. 825–834, 2007.

[3] M. Pucher, D. Schabus, and J. Yamagishi, "Synthesis of fast speech with interpolation of adapted HSMMs and its evaluation by blind and sighted listeners," in *Proc. Interspeech*, Chiba, Japan, Sept. 2010, pp. 2186–2189.

[4] T. Gay, "Effect of speaking rate on vowel formant movements," *J. Acoust. Soc. Am.*, vol. 63, no. 1, pp. 223–230, 1978.

[5] E. Janse, S. Nooteboom, and H. Quené, "Word-level intelligibility of time-compressed speech: Prosodic and segmental factors," *Speech Comm.*, vol. 41, no. 2, pp. 287–301, 2003.

[6] R. F. Port, "Linguistic timing factors in combination," *J. Acoust. Soc. Am.*, vol. 69, no. 1, pp. 262–274, 1981.

[7] F. Goldman-Eisler, *Psycholinguistics: Experiments in spontaneous speech*. London: Academic Press, 1968.

[8] R. Greisbach, "Reading aloud at maximal speed," *Speech Comm.*, vol. 11, no. 4-5, pp. 469 – 473, 1992.

[9] A. A. Sanderman and R. Collier, "Prosodic phrasing and comprehension," *Language and Speech*, vol. 40, no. 4, pp. 391–409, 1997.

[10] E. Janse, "Word perception in fast speech: artificially time-compressed vs. naturally produced fast speech," *Speech Comm.*, vol. 42, no. 2, pp. 155–173, 2004.

[11] M. Covell, M. Withgott, and M. Slaney, "Mach1: Nonuniform time-scale modification of speech," in *Proc. ICASSP*, vol. 1. Seattle, USA: IEEE, May 1998, pp. 349–352.

[12] L. He and A. Gupta, "Exploring benefits of non-linear time compression," in *Proc. ACM Int. Conf. on Multimedia*. Ottawa, Canada: ACM, Sept. 2001, pp. 382–391.

[13] J. Lebeter and S. Saunders, "The effects of time compression on the comprehension of natural and synthetic speech," *Working Papers of the Linguistics Circle*, vol. 20, no. 1, pp. 63–81, 2010.

[14] A. K. Syrdal, H. T. Bunnell, S. R. Hertz, T. Mishra, M. F. Spiegel, C. Bickley, D. Rekart, and M. J. Makashay, "Text-to-speech intelligibility across speech rates." in *Proc. Interspeech*, Portland, USA, Sept. 2012.

[15] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Duration modeling for HMM-based speech synthesis," in *Proc. ICSLP*, Sydney, Australia, Dec. 1998, pp. 29–32.

[16] T. Yoshimura, T. Masuko, K. Tokuda, T. Kobayashi, and T. Kitamura, "Speaker interpolation for HMM-based speech synthesis system," *Acoustical Science and Technology*, vol. 21, no. 4, pp. 199–206, Jan. 2000.

[17] M. Tachibana, J. Yamagishi, T. Masuko, and T. Kobayashi, "Speech synthesis with various emotional expressions and speaking styles by style interpolation and morphing," *IEICE Trans. Inf. Syst.*, vol. E88-D, no. 11, pp. 2484–2491, Nov. 2005.

[18] W. Verhelst and M. Roelands, "An overlap-add technique based on waveform similarity (WSOLA) for high quality time-scale modification of speech," in *Proc. ICASSP*, vol. 2, April 1993, pp. 554–557.

[19] T. Fukada, K. Tokuda, T. Kobayashi, and S. Imai, "An adaptive algorithm for mel-cepstral analysis of speech," in *Proc. ICASSP*, vol. 1, San Francisco, USA, March 1992, pp. 137–140.

[20] H. Kawahara, I. Masuda-Katsuse, and A. 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 Comm.*, vol. 27, pp. 187–207, 1999.

[21] J. Yamagishi, T. Kobayashi, Y. Nakano, K. Ogata, and J. Isogai, "Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 17, no. 1, pp. 66 –83, 2009.

[22] L. Rabiner and R. Schafer, *Theory and Applications of Digital Speech Processing*, 1st ed. Upper Saddle River, NJ, USA: Prentice Hall Press, 2010.

[23] IEEE, "IEEE recommended practice for speech quality measurement," *IEEE Trans. on Audio and Electroacoustics*, vol. 17, no. 3, pp. 225 – 246, 1969.