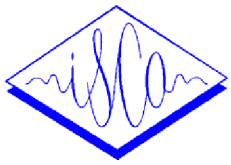

7th ISCA Workshop on
Speech Synthesis (SSW-7)
Kyoto, Japan
September 22-24, 2010

Comparison of Formant Enhancement Methods for HMM-Based Speech Synthesis

Tuomo Raitio[1], *Antti Suni*[2], *Hannu Pulakka*[1],
Martti Vainio[2], *Paavo Alku*[1]

[1]Department of Signal Processing and Acoustics, Aalto University, Helsinki, Finland
[2]Department of Speech Sciences, University of Helsinki, Helsinki, Finland
tuomo.raitio@tkk.fi

Abstract

Hidden Markov model (HMM) based speech synthesis has a tendency to over-smooth the spectral envelope of speech, which makes the speech sound muffled. One means to compensate for the over-smoothing is to enhance the formants of the spectral model. This paper compares the performance of different formant enhancement methods, and studies the enhancement of the formants prior to HMM training in order to preemptively compensate for the over-smoothing. A new method for enhancing the formants of an all-pole model is also introduced. Experiments indicate that the formant enhancement prior to HMM training improves the quality of synthetic speech by providing sharper formants, and the performance of the new formant enhancement method is similar to the existing method.

Index Terms: speech synthesis, hidden Markov model, over-smoothing, formant enhancement

1. Introduction

Hidden Markov model (HMM) based parametric speech synthesis techniques [1, 2, 3, 4] have gained much popularity due to their flexibility, but the quality and naturalness of the parametric HMM-based speech synthesis has remained poorer than that of the best unit selection methods. This degradation is mainly caused by three factors: oversimplified vocoder techniques, acoustic modeling accuracy, and over-smoothing of the generated speech parameters [4]. This paper concentrates on alleviating the third factor, the over-smoothing.

Over-smoothing is inherent to statistical parametric speech synthesis, since the speech sounds are synthesized from a parameter set that is an average of similarly sounding instances of natural speech. The averaging occurs at least in two phases. First, the training process of HMMs statistically averages the speech parameters in order to construct robust models. Second, in the synthesis stage, the parameter generation algorithm uses dynamic features [1] that enable the generation of continuous and smooth parameter trajectories. Although smoothing reduces artefacts caused by rapid changes in the speech parameters, the generated parameters are often over-smoothed, which results in muffled voice quality.

The over-smoothing is especially severe perceptually when it affects the formant structure of speech. There are two common methods in speech synthesis for modeling the speech spectrum: linear predictive coding (LPC) [5] and mel-cepstral analysis and synthesis [6], both of which are prone to the over-smoothing effect caused by the statistical modeling and parameter generation. The over-smoothing results in formants of large bandwidth (low Q-value), which deteriorates especially the perceptability of vowels.

The over-smoothing can be alleviated by using better acoustical modeling, such as trajectory HMMs [7] or minimum generation error (MGE) training [8]. However, the over-smoothing cannot be completely prevented with acoustic modeling. Several previous studies have addressed the problem of over-smoothing in the statistical context by using, e.g., global variance [9], or reducing the over-smoothing by explicitly using original speech material for generating the parameters [10]. However, the most straightforward way to compensate for the over-smoothing is to emphasize the spectral envelope by post-filtering [11]. Post-filtering modifies the spectral model so that dynamics between the formant peaks and the spectral valleys is increased, aiming at a more prominent formant structure. Post-filtering was originally developed for speech coding, but it is widely adopted to speech synthesis for enhancing the spectral structure of the synthesis filter. However, extensive use of post-filtering results in an overly sharp formant structure, which makes the quality of the synthetic speech unnatural.

In this paper, different *formant enhancement methods* are studied in order to tackle problems caused by over-smoothing. First, a comparison between different formant enhancement methods is carried out. A new method for enhancing the all-pole spectral model is introduced and compared to a widely used existing method. Second, the effect of applying the spectral enhancement before the parameter training is assessed. Usually, the averaging effect of the statistical modeling is compensated for after the speech parameter generation. In contrast, this paper also studies the enhancement of the formants prior to HMM training in order to preemptively compensate for the over-smoothing.

The rest of the paper is organized as follows: Section 2 describes two all-pole model based formant enhancement methods, one existing method and one completely new method. The idea of pre-enhancement is also introduced. Objective experiments and experiments on using the methods in a HMM-based speech synthesis framework are described in Section 3. Discussion on the proposed methods is presented in Section 4, and Section 5 concludes the paper.

2. Formant Enhancement Methods

Formant enhancement methods were originally developed for speech codecs in order to alleviate the perceptual effect of the quantization noise. The output of the decoder was processed with an adaptive post-filter [12] that emphasized the formants and attenuated the noise at spectral valleys. The post-filter in

[12] consists of three main parts: short-term post-filter, long-term post-filter, and spectral tilt compensation filter. The short-term part of the post-filter has a pole-zero structure, and it is derived from the LPC filter by scaling the radii of the poles and zeros.

In speech synthesis, similar approaches can be used to enhance the formant structure over-smoothed by the statistical mapping. There are two main methods for enhancing the structure of the synthesis filter. First, for mel-cepstrum based modeling, post-filtering is widely used, especially as the the post-filter is implemented in the HTS system [13, 14], a commonly used platform for HMM-based speech synthesis. Recently, a line spectral frequency (LSF) [17] based post-filtering method was introduced in [11] for enhancing the all-pole spectrum. These two post-filtering methods are based on different spectral modeling techniques, and this paper will focus on methods based on the all-pole modeling. The LSF-based enhancement method will be used as a baseline system for the spectral enhancement, and it is described more detailed in the following section. A new spectral enhancement method based on the modification of the all-pole power spectrum is introduced in Section 2.2.

2.1. LSF-based Formant Enhancement Method

The post-filter method described in [11] is based on converting the all-pole model to line spectral frequencies (LSF) and modifying the LSF positions in order to make the spectral peaks sharper. The modified LSFs are calculated from index two to $D-1$ recursively as

$$l'_i = l_{i-1} + d_{i-1} + \frac{d_{i-1}^2}{d_{i-1}^2 + d_i^2}\big((l_{i+1} - l_{i-1}) - (d_i + d_{i-1})\big) \quad (1)$$

$$d_i = \alpha(l_{i+1} - l_i), \quad 0 < \alpha < 1, \quad i = 2, \ldots, D-1 \quad (2)$$

[11] where l_i is the original ith LSF of a frame, l'_i is the modified ith LSF, D is the order of the all-pole model, and α controls the degree of the enhancement. The effect of the enhancement increases as α approaches zero. The algorithm shifts the LSFs that are close to each other even closer, which makes the spectral peaks sharper. This effectively enhances the formant structure, but it also makes the peaks slightly shift from the original positions. Figure 1 illustrates the enhancement procedure showing the original and the enhanced spectra and the corresponding LSF positions.

The LSF-based post-filtering method, described in Equations 1 and 2, was shortly covered in the paper by Ling et al. [11], but the rationales behind the method are not documented in detail. Studies of the method are also published in Chinese [15] and Japanese [16].

2.2. Proposed Formant Enhancement Method

A new technique, referred to as the LPC-based formant enhancement method, for enhancing the structure of speech spectrum is proposed. The method is based on modifying the power spectrum of the LPC model and then re-evaluating LPC based on the modified power spectrum.

The enhancement begins with the evaluation of the power spectrum from the LPC coefficients. This yields the spectral model of speech, which can then be modified in order to emphasize certain spectral components. The power spectrum is modified so that the low-energy parts of the spectrum, i.e., the valleys, are additionally reduced by multiplying spectral components in those regions with a small real-valued coefficient. The spectral peaks are left unmodified. All the spectral peaks are easily found from the smooth LPC envelope by searching for the zero-crossing points in the differentiated spectral envelope. There are two parameters in the modification procedure that control the effect of the enhancement: the width (δ) of the unmodified area withing a spectral peak, and the coefficient γ ($0 \leq \gamma < 1$) that reduces the low-energy areas of the spectrum. These parameters are illustrated in Figure 2.

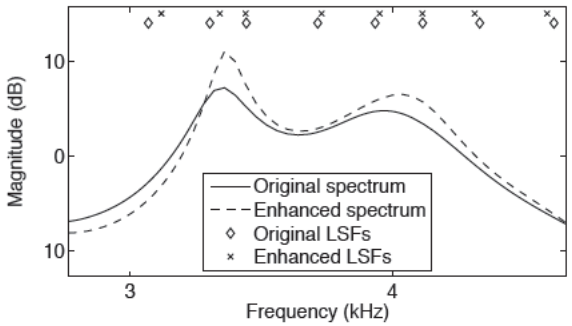

Figure 1: *Illustration of the LSF-based formant enhancement method. The original (solid) and enhanced (dashed) spectra are shown, and LSF positions in frequency are indicated for original ($\diamond$) and enhanced ($\times$) spectra.*

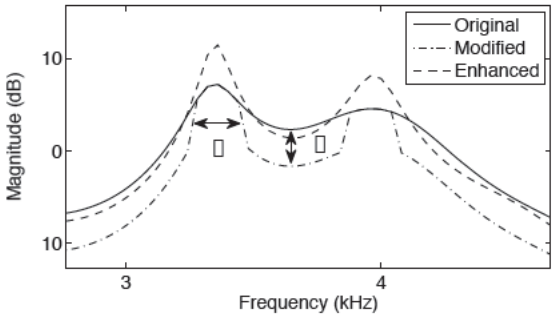

Figure 2: *Parameters of the LPC-based enhancement method that control the effect of the enhancement. Parameter δ controls the width of the unmodified area within a spectral peak, and parameter γ ($0 \leq \gamma < 1$) controls the reduction of the low-energy areas of the spectrum. The original power spectrum (solid), the modified power spectrum (dash-dotted), and the enhanced spectrum (dashed) are shown.*

The modified power spectrum is inverse Fourier transformed to get a new autocorrelation function. This autocorrelation function is used in the Yule–Walker equations [5] to compute a new LPC filter. Since LPC analysis focuses in the frequency domain on spectral peaks, the new LPC model will most likely show sharper spectral resonances than the original LPC filter. Figure 3 shows an example of the proposed formant enhancement procedure.

Although the modification is symmetric with respect to the spectral peaks, the different proportions of energy on each side of a peak can make the modification to slightly shift the spectral peaks. Formant shift due to the spectral enhancement will be assessed in Section 3.

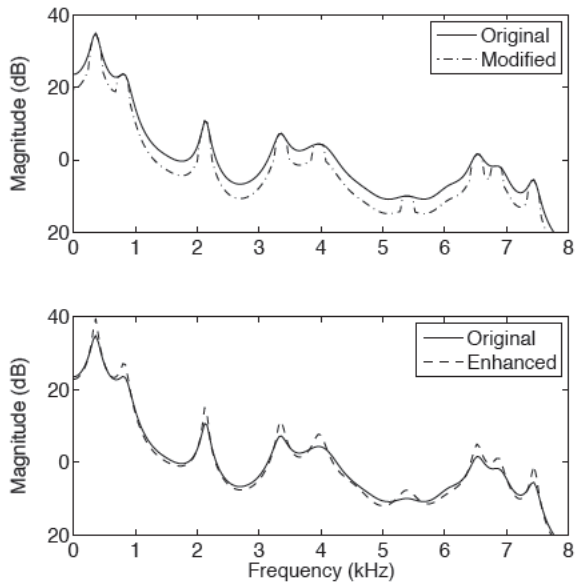

Figure 3: *Illustration of the proposed formant enhancement method. In the upper figure, the original (solid) and the modified (dashed) power spectra are shown. The energy of the modified spectrum is reduced in the low-energy regions, which results in relatively higher energy in spectral peaks. In the lower figure, the original (solid) and the enhanced (dashed) LPC spectral envelopes are shown. The enhanced spectrum is calculated from the modified power spectrum, and thus the spectral peaks are enhanced.*

2.3. Formant Enhancement Prior to HMM Training

Conventionally, the averaging effect of the statistical modeling is compensated for after the speech parameter generation. In contrast, this paper studies the enhancement of formants prior to HMM training in order to preemptively compensate for the over-smoothing. The flow chart in Figure 4 shows two possible formant enhancement implementations: (1) enhancement prior to the training of HMMs (pre-enhancement), and (2) enhancement after the parameter generation from HMMs (post-filter).

In the case that the enhancement is performed before the training stage, the spectrum that is used to train the HMMs has more dynamics compared to the conventional LPC spectrum. In the synthesis stage, however, the generated parameters are slightly smoothed, and they should correspond more closely to the real speech spectrum if the pre-enhancement parameters are chosen appropriately. As a result, the over-smoothing of the formants is effectively prevented. Additionally, the pre-enhancement provides formant information that has higher dynamics for training material, which also affects the entire training process of the HMMs. More prominent formant information may yield more robust models and enhance the quality of synthesized speech. However, the training and parameter generation are complex processes, and thus the effect of pre-enhancement on the overall speech quality is difficult to predict.

Figure 5 shows two synthetic speech spectra of which the left one is obtained without formant enhancement, and the right one is from a system that uses formant pre-enhancement before the training of HMMs. The spectral peaks are sharper in the spectrum generated by the enhanced system.

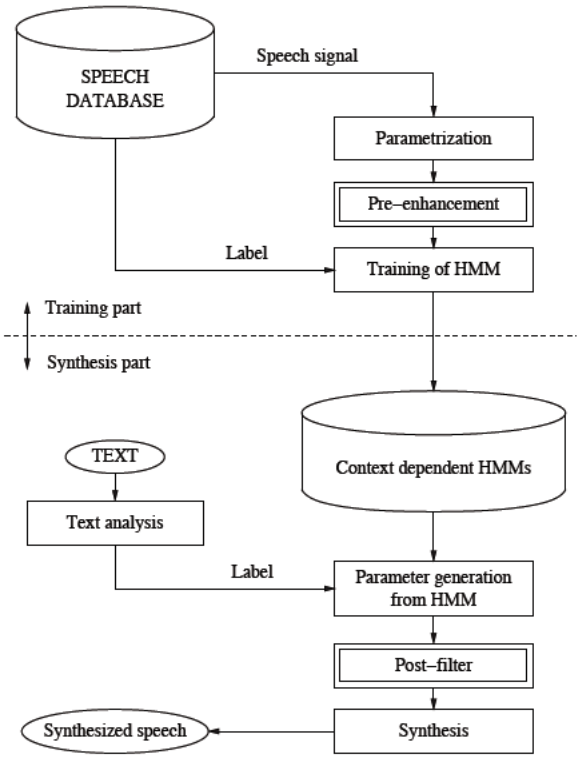

Figure 4: *Overview of a speech synthesis system with two options for compensating for the over-smoothing effect. First, the over-smoothing can be preemptively compensated for before the HMM training (pre-enhancement), or second, the compensation can be performed after the parameter generation (post-filter).*

3. Experiments

The formant enhancement methods were evaluated both objectively by measuring the emphasis of individual formants, and subjectively by assessing the quality of synthetic speech with different formant enhancement methods.

3.1. Objective Evaluation

The performance of the two formant enhancement methods, LSF-based (Section 2.1) and LPC-based (Section 2.2) methods, were first evaluated by analyzing their effect on formants. A database of eight Finnish vowels [a, æ, e, i, o, œ, u, y] spoken by ten Finnish speaker (5 males and 5 females) was used. Vowels were first analyzed with LPC with a fixed first-order pre-emphasis (zero at $z = 0.68$), and then formant enhancement methods were applied to the LP coefficients. The enhancement performance was measured by comparing the bandwidths (-3 dB) of the two first formants of every vowel before and after the enhancement. The ratio of the enhanced and the original bandwidth ($R = B_{\text{enh}}/B_{\text{orig}}$) was evaluated, indicating the sharpening of the formants. The formant shift due to the enhancement was also measured, and was indicated in percents from the original formant center frequency. The LSF-based formant enhancement method was evaluated on three different values of α (0.3, 0.4, and 0.5), and the LPC-based formant enhancement method was evaluated on three values of γ (0.2, 0.3, and 0.4). An appropriate range of the parameters that control the effect of the enhancement (α and γ) was selected according to the rea-

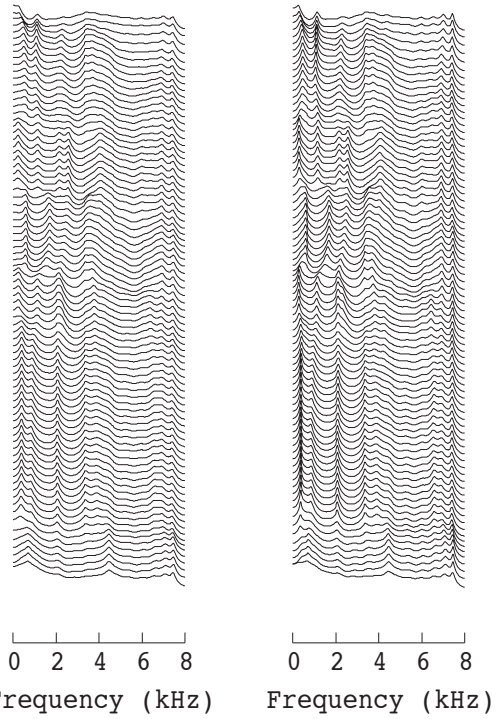

Figure 5: *Illustration of the formant pre-enhancement using the proposed LPC-based method. Left figure shows the spectral envelope of synthesized speech without enhancement. Right figure shows the spectral envelope of synthetic speech generated by a system using the spectral enhancement before HMM training.*

sonable functioning of each method. The parameter δ for the LPC-based method was set to 160 Hz on the grounds of small-scale experiments. The results of the analytical evaluation are shown in Table 1.

The results show that both methods effectively emphasize formants, and the amount of emphasis can be controlled by their parameters. The LPC-based method tends to shift the first formant slightly more than the LSF-based method, whereas the LSF-based method tends to shift the second formant more than the LPC-based method. Flanagan [18] has reported that just-noticeable differences (JND) for vowel formant frequencies of F1 and F2 vary from three to five percent of the formant frequency. Some of the formant shifts shown in Table 1 are above the JND. However, the highest average shift of F1 for the LPC-based method, almost 7 %, is obtained with a high degree of enhancement ($R = 0.28$).

3.2. Subjective Evaluation with a HMM-Based Synthesizer

The formant enhancement methods were evaluated subjectively by assessing the quality of synthetic speech with different formant enhancement setups. A total of four systems were compared:

1. No formant enhancement
2. LPC-based pre-enhancement ($\gamma = 0.3$, $\delta = 160$)
3. LSF-based pre-enhancement ($\alpha = 0.4$)
4. Post-filtering with LSF-based enhancement ($\alpha = 0.5$)

The parameters α and γ that control the effect of formant enhancement were chosen by constructing each system with several parameter values and selecting the best sounding one by expert listeners. The value of δ for the LPC-based enhancement was selected according to previous experiments. The effect of the formant enhancement was assessed to be equal in all the systems within the test, except for the one without formant enhancement.

Table 1: *Average bandwidth (−3 dB) ratio* ($R = B_{\text{enh}}/B_{\text{orig}}$) *and the average formant shift* (ΔF) *of the first two formants for the two methods. The number after the method indicates the values of* α *and* γ *for the LSF and LPC-based methods, respectively.*

	Method	R	ΔF (%)
F1	LSF-Enh-03	0.46	3.38
	LSF-Enh-04	0.52	2.82
	LSF-Enh-05	0.58	2.28
	LPC-Enh-02	0.28	6.95
	LPC-Enh-03	0.37	5.83
	LPC-Enh-04	0.45	4.79
F2	LSF-Enh-03	0.41	1.96
	LSF-Enh-04	0.48	1.73
	LSF-Enh-05	0.55	1.49
	LPC-Enh-02	0.34	0.55
	LPC-Enh-03	0.42	0.49
	LPC-Enh-04	0.51	0.42

A prosodically annotated database of 600 phonetically rich sentences spoken by a 39-year-old Finnish male speaker, comprising one hour of speech material was used to train the synthesizer. Speech was sampled at 16 kHz.

3.2.1. Speech Synthesis System

An HMM-based speech synthesizer [19, 20] that utilizes glottal inverse filtering for separating the vocal tract from the glottal source was used as a test system. Although the synthesizer is built on a basic framework of an HMM-based speech synthesis system [13], the parametrization and synthesis methods are different from other HMM-based synthesizers, and therefore they are explained in detail below.

In the parametrization, the signal is first high-pass filtered and windowed with a rectangular window to 25-ms frames at 5-ms intervals. The speech features, presented in Table 2, are then extracted from each frame. The log-energy of the window is evaluated, after which glottal inverse filtering is performed in order to estimate the glottal volume velocity waveform from the speech signal. An automatic inverse filtering method, Iterative Adaptive Inverse Filtering (IAIF) [22, 21], is utilized in the system. IAIF iteratively cancels the effects of the vocal tract and the lip radiation from the speech signal using all-pole modeling. The outputs of the IAIF block are the estimated glottal flow signal and the LPC model of the vocal tract. The spectral envelope of the glottal flow is further parametrized with LPC. The fundamental frequency and a harmonic-to-noise ratio (HNR) are determined from the glottal flow signal. The HNR values are based on the ratio between the upper and lower smoothed spectral envelopes (defined by the harmonic peaks and interharmonic valleys, respectively) and averaged across five frequency bands according to the equivalent rectangular bandwidth (ERB) scale [23]. LPC models of the vocal tract and the voice source are further converted to LSFs [17]. In case

of unvoiced speech, conventional LPC is used to evaluate the spectral model of speech.

For HMM training, the speech data was labeled with a rich set of phonologically relevant contextual features [25]. Each parameter type was assigned to its own stream and clustered separately. Otherwise, the HMM-training followed the default HTS 2.1 procedure. Global variance [9] was not used.

In the synthesis part, the excitation signal consists of voiced and unvoiced sound sources. The basis of the voiced sound source is a glottal flow pulse extracted from a natural vowel. By interpolating the real glottal flow pulse according to F_0 and scaling in magnitude according to the energy measure, a pulse train comprising a series of individual glottal flow pulses is generated. The amount of noise in the excitation is matched by manipulating the phase and magnitude of the spectrum of each pulse according to the harmonic-to-noise measure. Furthermore, the spectral tilt of each pulse is modified according to the all-pole spectrum generated by the HMM. This is achieved by filtering the pulse train with an adaptive IIR filter which flattens the spectrum of the pulse train and applies the desired spectrum. For voiced excitation, the lip radiation effect is modeled as a first-order differentiation operation. The unvoiced excitation is composed of white noise, whose gain is determined according to the energy measure generated by the HMM system. The LSFs are then interpolated and converted to LPC coefficients, and used for filtering the excitation signal.

3.2.2. Listening Tests

Two subjective listening tests were conducted to evaluate the quality of the four different setups of spectral enhancement. In both tests, a comparison category rating (CCR) test similar to [26] was used to assess the quality of synthetic speech samples. In the CCR test, the listeners were presented with a pair of speech samples on each trial, and they were asked to assess the quality of the second sample compared to the first one on the comparison mean opinion score (CMOS) scale. Eleven Finnish listeners compared a total of 70 speech sample pairs in each test. The ranking of the methods was evaluated by averaging the scores of the CCR test for each method.

In the first test, the overall performance of each method was evaluated. Ten randomly chosen sentences from the held-out data were used for generating the test samples for each method. The ranking of the four methods with 95 % confidence intervals is shown in Fig. 6.

In the second test, synthetic speech samples were generated from aligned labels and predefined F_0 vectors in order to assess only the quality of the formant enhancement. Thus, different training data due to the pre-enhancement did not have effect on the durations and the F_0 values. Ten sentences, different from the ones used in the first test, were used for generating the test samples for each method. The ranking of the four methods with aligned data with 95 % confidence intervals is shown in Fig. 7.

Table 2: Speech features and the number of parameters.

Feature	Parameters per frame
Fundamental frequency	1
Energy	1
Harmonic-to-noise ratio	5
Voice source spectrum	10
Vocal tract spectrum	30

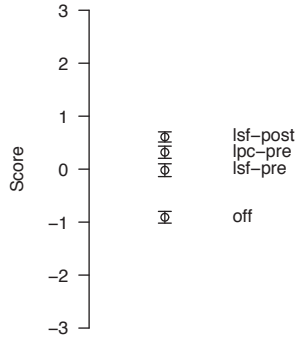

Figure 6: *Ranking of the four systems with synthetic speech: No formant enhancement (off), formant enhancement prior to HMM training with LSF-based method (lsf-pre) and LPC-based method (lpc-pre), and post-enhancement with LSF-based method (lsf-post). The mean score has no explicit meaning, but the distances between the scores are essential. The 95 % confidence intervals are presented for each score.*

The results of the first test show that all the spectral enhancement methods greatly improve the quality of speech compared to unenhanced speech. The system using post-enhancement with LSF-based method (lsf-post) was graded best, but the two systems with pre-enhancement (lpc-pre and lsf-pre) are close to the best one.

The results of the second test with aligned synthetic speech with pre-defined F_0 values show that LSF-based post-enhancement (lsf-post) and LPC-based pre-enhancement (lpc-pre) were graded equally good. The system using LSF-based pre-enhancement (lsf-pre) was graded slightly worse than the two best systems. The unenhanced system (off) was graded clearly worse than any of the enhanced systems.

The differences between the first and the second test show that the pre-enhancement methods have effect on the overall quality of the system. The prosodic features, durations and F_0, are slightly degraded with the systems using pre-enhancement.

4. Discussion

The results show that the spectral enhancement prior to HMM training effectively alleviates the over-smoothing. The new method (lpc-pre) performed better than the existing method (lsf-pre) in pre-enhancement and produced equally good results when compared to the existing method when it was used in post-processing (lsf-post). However, comparison between the two listening tests, not-aligned and aligned one, shows that the pre-enhancement slightly degrades the modeling of durations and F_0. This is somewhat surprising since more prominent formant information was assumed to produce more robust statistical models. The rationales for this call for more studies.

This paper provided preliminary results on the proposed formant enhancement method. However, there are several ways to modify the method. For example, the parameter γ may be defined as a varying vector ranging from zero to sampling frequency. Since the first two formants are most important in the perception of vowels, the enhancement could be configured to emphasize the lowest frequencies (e.g., 0–3500 Hz) more than the highest frequencies. Similarly, the parameter δ could be

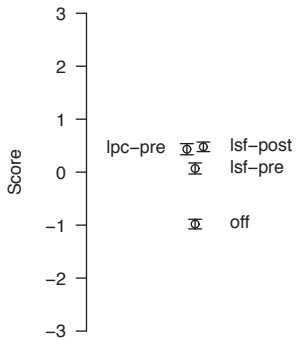

Figure 7: *Ranking of the four systems with synthetic speech with normalized durations and F_0 values: No formant enhancement (off), formant enhancement prior to HMM training with LSF-based method (lsf-pre) and LPC-based method (lpc-pre), and post-enhancement with LSF-based method (lsf-post). The mean score has no explicit meaning, but the distances between the scores are essential. The 95 % confidence intervals are presented for each score.*

varied according the frequency, for example to reduce the occasional shifting of the first formant. In addition, information from adjacent frames could be used to improve the robustness of the enhancement.

5. Conclusions

In this study, a comparison between different formant enhancement methods for alleviating the over-smoothing of the statistical mapping in HMM-based speech synthesis was conducted. A new method for formant enhancement was introduced, and the pre-enhancement for preemptively compensating for the over-smoothing was experimented. The results showed that the new method performed similarly or better to the existing method, and the pre-enhancement produced almost as good quality as the post-enhancement. Although the results are promising, more experiments are required to fully evaluate the performance of the pre-enhancement and the new formant enhancement method.

6. Acknowledgements

This project is supported by Nokia, the Academy of Finland (projects 135003, 107606, 1128204, 1218259, research programme LASTU), and MIDE UI-ART.

7. References

[1] Tokuda, K., Masuko, T., Yamada, T., Kobayashi, T. and Imai, S., "An algorithm for speech parameter generation from continuous mixture HMMs with dynamic features", in Proc. Eurospeech, 1:757–760, 1995.

[2] Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T. and Kitamura, T., "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis", in Proc. Eurospeech, 2374–2350, Sep. 1999.

[3] Tokuda, K., Zen, H. and Black, A. W., "An HMM-based speech synthesis system applied to English", in Proc. 2002 IEEE Workshop on Speech Synthesis, 227–230, Sep. 2002.

[4] Zen, H., Tokuda, K. and Black, A. W., "Statistical parametric speech synthesis", Speech Commun., 51(11):1039–1064, 2009.

[5] Makhoul, J., "Linear prediction: A tutorial review", in Proc. of the IEEE, 63(4):561–580, Apr. 1975.

[6] Fukada, T., Tokuda, K., Kobayashi, T., Imai, S., "An adaptive algorithm for mel-cepstral analysis of speech", in Proc. ICASSP, 137–140, 1992.

[7] Tokuda, K., Zen, H. and Kitamura, T., "Trajectory modeling based on HMMs with the explicit relationship between static and dynamic features", In Proc. Eurospeech, 865–868. Sep. 2003.

[8] Wu, Y.-J. and Wang, R.-H., "Minimum generation error training for HMM-based speech synthesis", in Proc. ICASSP, 89–92, 2006.

[9] Toda, T. and Tokuda, K., "A speech parameter generation algorithm considering global variance for HMM-based speech synthesis", IEICE Trans. Inf. & Syst., E90-D(5):816–824, May 2007.

[10] Masuko, T., Tokuda, K. and Kobayashi, T., "A study on conditional parameter generation from HMM based on maximum likelihood criterion", in Proc. Autumn Meeting of ASJ, 209–210, 2003. [in Japanese]

[11] Ling, Z.-H., Wu Y.-J., Wang Y.-P., Qin, L. and Wang, R.-H., "USTC system for Blizzard Challenge 2006 an improved HMM-based speech synthesis method", Blizzard Challenge Workshop, 2006.

[12] Chen, J.-H. and Gersho, A., "Adaptive postfiltering for quality enhancement of coded speech", IEEE Trans. on Speech and Audio Processing, 3(1):59–71, Jan. 1995.

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

[14] HTS, "HMM-based speech synthesis system", Apr. 2009. Online: http://hts.sp.nitech.ac.jp

[15] Wu, Y.-J., "Research on HMM-based Speech Synthesis", Ph.D Thesis, University of Science and Technology of China, 2006. [in Chinese]

[16] Oura, K., Zen, H., Nankaku, Y., Lee, A. and Tokuda, K., "Postfiltering for HMM-based speech synthesis using mel-LSPs", Proc. Autumn Meeting of ASJ, pp. 367–368, 2007. [in Japanese]

[17] Soong, F. K. and Juang, B.-H., "Line spectrum pair (LSP) and speech data compression", Proc. ICASSP, 9:37–40, 1984.

[18] Flanagan, J. L., "Speech Analysis, Synthesis and Perception", 2nd ed., Springer-Verlag, 1972.

[19] Raitio, T., Suni, A., Pulakka, H., Vainio, M. and Alku, P., "HMM-based Finnish text-to-speech system utilizing glottal inverse filtering", Proc. Interspeech, 2008.

[20] Raitio, T., Suni, A., Yamagishi, J., Pulakka, H., Nurminen, J., Vainio, M. and Alku, P., "HMM-based speech synthesis utilizing glottal inverse filtering", IEEE Trans. Audio, Speech, and Language Processing, (in press)

[21] Alku, P., Tiitinen, H. and Näätänen, R., "A method for generating natural-sounding speech stimuli for cognitive brain research", Clinical Neurophysiology, 110:1329–1333, 1999.

[22] Alku, P., "Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering", Speech Commun. 11(2–3):109–118, Jun. 1992.

[23] Moore, B. C. J. and Glasberg, B. R., "A revision of Zwicker's loudness model", ACTA Acustica, 82:335–345, 1996.

[24] Zen, H., Tokuda, K., Masuko, T., Kobayashi, T. and Kitamura, T., "Hidden semi-Markov model based speech synthesis", Proc. Interspeech, 2:1397–1400, Oct. 2004.

[25] Vainio, M., Suni, A. and Sirjola, P., "Accent and prominence in Finnish speech synthesis", Proc. of the 10th International Conference on Speech and Computer, 309–312, Oct. 2005.

[26] Recommendation ITU-T P.800 "Methods for subjective determination of transmission quality", International Telecommunication Union, Aug. 1996.