Wideband Parametric Speech Synthesis Using Warped Linear Prediction

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

[1]Department of Signal Processing and Acoustics, Aalto University, Espoo, Finland
[2]Department of Behavioural Sciences, University of Helsinki, Helsinki, Finland
tuomo.raitio@aalto.fi, antti.suni@helsinki.fi

Abstract

This paper studies the use of warped linear prediction (WLP) for wideband parametric speech synthesis. As the sampling frequency is increased from the usual 16 kHz, linear frequency resolution of conventional linear prediction (LP) cannot efficiently model the speech spectrum. By using frequency warping that weights perceptually the most important formant information, spectral models with better accuracy and lower model orders can be utilized. In this work, WLP is embedded in a parametric speech synthesizer to efficiently create wideband synthetic speech. Experiments show that WLP-based wideband synthetic speech is rated better compared to narrowband speech and wideband LP-based speech.

Index Terms: statistical parametric speech synthesis, wideband, warped linear prediction, WLP

1. Introduction

Conventionally, parametric speech synthesizers utilize a sampling frequency of 16 kHz [1]. This corresponds to using an 8 kHz audio bandwidth which is sufficient to create rather pleasant and intelligible synthetic speech. However, speech sampled with 16 kHz still sounds slightly muffled compared to using higher rates. With increasing computational power and more advanced techniques available today, adopting higher sampling rates is a potential way to improve the quality of synthetic speech.

There are only a few previous studies where higher sampling rates have been used in parametric speech synthesis. Yamagishi and King [2] achieved enhanced feature extraction and improved speaker similarity at higher sampling rates. Similar improvements have also been reported by Stan et al. [3] who found that using speech sampled at 32 kHz or more resulted in better speaker similarity compared 16 kHz speech. However, they did not observe any improvements in naturalness or intelligibility. In both studies, mel-cepstral [4] type vocoders were used for feature extraction.

Although mel-cepstrum based spectral modeling techniques are prevalent in parametric speech synthesis, linear prediction (LP) based methods have provided similar performance [5, 6]. One of the fundamental differences between the two methods is the frequency resolution: mel-cepstrum utilizes a non-linear frequency resolution according to the Mel scale [7] while LP uses linear resolution. This difference has an important effect on the selection of the model order and thereby on the synthesis quality.

In LP, the order of the all-pole model is defined as the sampling frequency in kHz added by a small integer [8]. This selection of the prediction order enables computing all-pole filters capable of modeling the main formants of speech digitized with the corresponding sampling rate as well as the overall spectral tilt of the signal caused by the glottal excitation. Thus, as an example, for speech sampled at 16 kHz, a 20th order all-pole model would be a good choice. In parametric speech synthesis applications, however, it has been noted that slightly greater model orders improve the quality of synthetic speech [5]. Hence, a 30th order all-pole model might be better for parametric speech synthesis.

If speech bandwidth is increased, greater all-pole model orders are required. For 44.1 kHz speech, a model order of about 50 would be a reasonable choice. However, the number of actual resonances in natural speech does not increase linearly, and perceptually the most important formant information is at low frequencies. Thus, a linear frequency resolution embedded in conventional LP might not be justified if the bandwidth is greatly increased. In addition, it is computationally more expensive to train a large number of components in parametric speech synthesis, and the accuracy of a high-order all-pole model may suffer in estimation and statistical modeling.

In this paper, frequency warped linear prediction (WLP) is experimented with for efficiently creating high sampling rate synthetic speech. First, WLP is shortly introduced, after which the experimental results on using conventional LP and WLP in parametric speech synthesis are presented and discussed.

2. Warped linear prediction (WLP)

Linear prediction on a warped frequency scale was first proposed by Strube [9] in 1980. Since then, the properties of WLP have been widely studied and it has been shown that all conventional parametric spectral estimation and linear filtering methods can be warped in a straightforward way [10, 11]. The major advantage of WLP is that the frequency resolution can be made closer to that of human hearing. Thus, WLP leads to either perceptually more accurate spectral models or smaller model orders with equal accuracy.

In WLP, spectral representation is modified by replacing the unit delay elements by first-order all-pass filters:

$$D(z) = \frac{z^{-1} - \lambda}{1 - \lambda z^{-1}} \tag{1}$$

By definition, the magnitude response of the filter is constant, but the phase response of $D(z)$ defines the frequency mapping:

$$\tilde{\omega} = \text{arg}(D(e^{-i\omega})) = \omega + 2\text{arctan}\left(\frac{\lambda\text{sin}(\omega)}{1 - \lambda\text{cos}(\omega)}\right) \tag{2}$$

If $\lambda > 0$, the frequency resolution is better at low frequencies than high frequencies, as in human hearing. Fig. 1 illustrates warped frequency scales with different values of λ.

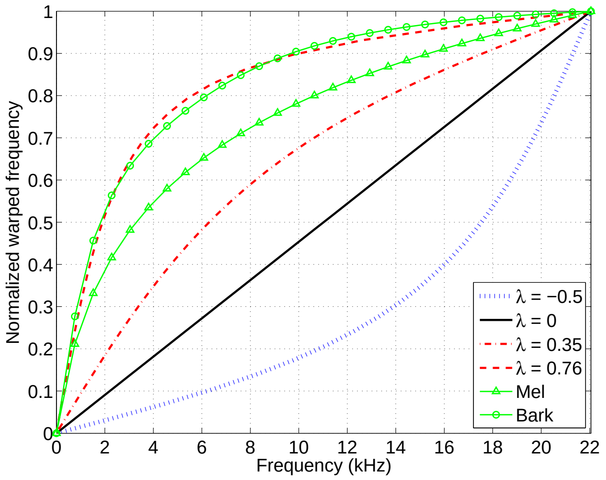

Figure 1: *Illustration of the frequency resolution of WLP with different values of λ. Mel and Bark scales are shown for reference. Steeper curve corresponds to higher frequency resolution.*

Computing WLP is identical to conventional LP except that the autocorrelation sequence is computed in a warped domain by replacing the unit delays with all-pass sections [10]. WLP coefficients can be then computed from the warped autocorrelation just like in conventional autocorrelation LP, e.g., by using the *Levinson-Durbin* algorithm.

Warped finite impulse response (FIR) filter can be directly realized by replacing the unit delays with all-pass sections:

$$A(z) = 1 - \sum_{k=1}^{N} a_k D(z)^k \qquad (3)$$

Warped infinite impulse response (IIR) filter leads to delay-free loops if directly realized by all-pass sections. Techniques for implementing warped IIR filters are presented in [12].

3. Using WLP in parametric speech synthesis

The benefits of WLP, better spectral modeling accuracy in relevant frequencies and smaller model order, are both desirable properties in parametric speech synthesis, especially with speech sampled at high rates. However, WLP has not been utilized in parametric speech synthesis.

WLP-based parametric speech synthesis is identical to LP-based synthesis except that all linear predictive filters, both FIRs and IIRs, are replaced with their warped counterparts. In this work, we have implemented warped spectral modeling methods to our LP-based parametric speech synthesis system, which is described below.

3.1. Parametric speech synthesis system

Our synthesizer GlottHMM [5, 6] is built on a basic framework of a hidden Markov model (HMM) based speech synthesis system [13], but it uses a specific type of vocoder for parameterizing and reconstructing speech. First, glottal inverse filtering [14] is used in order to decompose speech into a vocal tract filter and the voice source. The purpose of this decomposition is to accurately estimate and model both of the speech production components: source and filter. The voice source is parameterized with energy, fundamental frequency (F0), harmonic-to-noise ratio (HNR), and source LP spectrum. The vocal tract spectrum is parameterized with (warped) LP. The parameters used by the vocoder are depicted in Table 1.

For HMM training, all LP-based parameters are converted to line spectral frequencies (LSF). One stream is assigned for each parameter type except energy and vocal tract LSFs, which are trained together. Standard HTS 2.1 method [13] is used for training the voices.

In synthesis stage, voice source is reconstructed by using a glottal flow pulse extracted from natural speech [5]. The pulse is interpolated according to F0, scaled in magnitude, and concatenated to create the voiced excitation. The voice source spectrum is modified with an IIR filter to match the given spectral measure [5]. White noise is used as an excitation for unvoiced sections. Finally, voiced and unvoiced excitations are combined and filtered with the vocal tract filter.

3.2. Selecting model order and warping coefficient

Model order and warping coefficient both affect the accuracy of the WLP model. In speech coding, WLP have resulted in perceptually identical output with lower model orders or better quality with equal model orders [10]. Usually the warping coefficient in WLP speech coders is chosen so that the warped frequency scale is very close to the Bark scale [15]. In parametric speech synthesis, similar behavior is expected, but the selection of model order and warping coefficient is not as straightforward due to differences in speech coding and synthesis; in speech synthesis, the estimated filter and residual are not used as such, but a generated excitation signal is modified with a filter that is an output of statistical modeling. These complex processes make the estimation of optimal model order and warping coefficient more experimental in nature.

4. Experiments

In order to test wideband speech synthesis based on WLP, three separate listening tests were conducted. First, the effect of warping in wideband speech synthesis was evaluated. Second, an overall listening test was performed in order to test the effects of bandwidth, model order, and warping to speech quality. Finally a similarity test was performed to find out if wideband speech is rated more similar to the original speaker, as is indicated by previous studies [2, 3].

The listening tests were performed in quiet listening booths with high-quality headphones. A total of 12 listeners, native speakers of Finnish working in the field of acoustics or signal processing, participated in the tests.

4.1. Speech material

A subset of a new Finnish ‘Heini’ database was used, consisting of 500 phonetically rich sentences and 270 sentences of continuous non-fiction, read by a 27-year-old female, comprising alto-

Table 1: 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	30–50

gether 50 642 phone instances. The database was automatically annotated with word prominence labels and segmented with HTS. Full-context labels with quinphones, word prominence, and typical positional and quantitative features were used in training.

All synthetic speech samples were generated by the system described in Section 3.1, with the addition of WLP in warped systems. Training of the voices was performed twice: after the first pass, the MDL factor controlling the decision tree sizes of LSF streams was adjusted in order to roughly match the model complexity (number of leaf nodes) between voices. The HNR parameter was not trained or used in order to keep the narrowband and wideband systems as similar as possible.

4.2. Effect of warping in wideband speech synthesis

In order to evaluate the effect of warping in wideband parametric speech synthesis, and to roughly estimate optimal warping coefficient, the following three systems were built:

A. 44.1 kHz speech, 30th order LP
B. 44.1 kHz speech, 30th order WLP, $\lambda = 0.35$
C. 44.1 kHz speech, 30th order WLP, $\lambda = 0.76$ (Bark scale)

All systems are trained with 44.1 kHz speech and 30th order spectral model. System A uses conventional LP while systems B and C utilize WLP with different warping coefficients. System B is warped with $\lambda = 0.35$, which was found to produce the best quality in analysis-synthesis experiments, and system C is warped with $\lambda = 0.76$, corresponding approximately to the Bark scale [15].

A comparison category rating (CCR) test was used to evaluate the differences between the systems. In CCR test, a listener is presented with speech sample pairs, and the task of the listener is to rate the quality difference between the samples on a continuous comparison mean opinion score (CMOS) scale. The scale ranges from -3 to 3 with verbal descriptions of quality. CMOS ratings were finally averaged for each system resulting in a ranking of the systems. Ten randomly chosen sentences from the held-out data were used for generating the samples. Each listener compared a total of 70 speech sample pairs.

The results of the CCR test are shown in Fig. 2. The mean scores have no explicit meaning, but the distances between the scores define the relative difference in quality. The 30th order LP was rated much lower in quality compared to warped systems, indicating that conventional 30th order LP is not capable of modeling wideband formant structure in detail. Warped systems were rated higher, and system B with $\lambda = 0.35$ gave slightly better result than system C, and is therefore used in further experiments.

4.3. Effect of bandwidth, model order, and warping

In order to test the effects of bandwidth, model order, and warping, a second listening test was conducted with the following systems:

A. 16 kHz speech, 30th order LP (baseline)
B. 44.1 kHz speech, 30th order WLP ($\lambda = 0.35$)
C. 44.1 kHz speech, 50th order LP
D. 44.1 kHz speech, 50th order WLP ($\lambda = 0.35$)

System A uses speech sampled at 16 kHz and a suitable 30th order LP model. This baseline system has been shown to yield good results [5, 6]. The rest of the systems were trained with 44.1 kHz speech that covers the whole human hearing range.

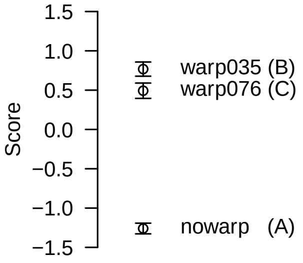

Figure 2: *Ranking of the systems showing the effect of warping for 44.1 kHz speech modeled by 30th order all-pole model. The mean score has no explicit meaning, but the distances between the scores are essential. The 95% confidence intervals are shown for each system.*

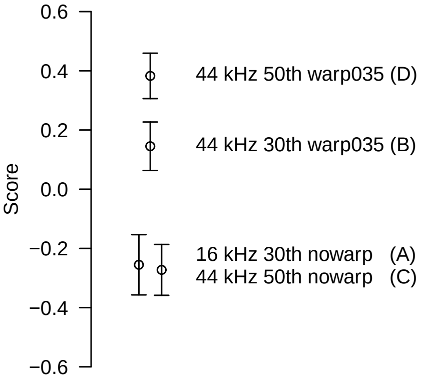

Figure 3: *Ranking of the systems showing the effect of bandwidth, model order, and warping. The mean score has no explicit meaning, but the distances between the scores are essential. The 95% confidence intervals are shown for each system.*

The second system, B, is the best system from the first listening test, i.e., the 30th order WLP model with $\lambda = 0.35$. Systems C and D have a model order of 50 in order to better model the whole audio range. In system C, conventional LP is used and in system D, WLP is used with the same $\lambda = 0.35$ as in the first test.

A CCR test was used again for evaluating the quality of the four systems. Ten randomly chosen sentences from the held-out data (different from the first test) were used. Each listener compared a total of 130 speech sample pairs.

The results of the second CCR test are shown in Fig. 3. The baseline system A with 16 kHz sampling rate and system C with 44.1 kHz sampling rate and 50th order LP show no statistically significant difference in quality. Systems B and D with 44.1 kHz speech and 30th and 50th order WLP were rated higher than conventional LP systems. System D with 50th order WLP gave the best result, still showing the positive effect of increased model order on quality. Results indicate the preference of WLP over conventional LP.

4.4. Similarity

The similarity of the speech samples to the original speaker was also evaluated. Systems A, B, C, and D described in the previ-

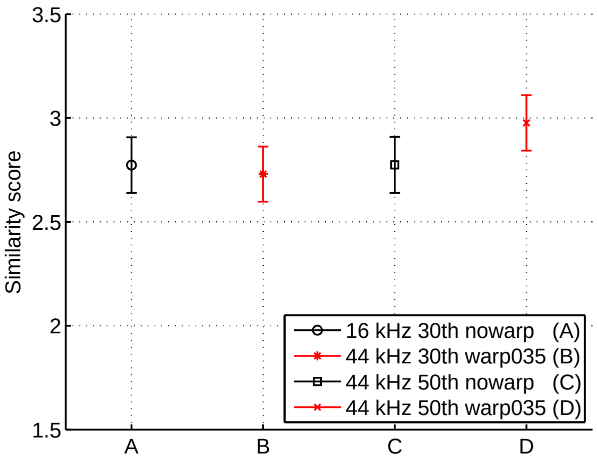

Figure 4: *Similarity scores for each systems with 95% confidence intervals.*

ous section were evaluated. In the similarity test, listeners were presented with four natural 44.1 kHz reference samples of the same speaker and one synthetic test sample. The task of the listener was to evaluate how similar the voice in the test sample sounded in comparison to the voice in the reference samples. A continuous scale was used ranging from 1 to 5 with the following verbal descriptions in the end points: *sounds like a totally different person* (1) and *sounds exactly the same person* (5). A similarity score was evaluated from the results for each system. Fifteen randomly chosen sentences from the held-out data (different from the first and the second test) were used. Each listener rated a total of 60 speech samples.

The similarity scores for each system are shown in Fig. 4. The results show no statistically significant differences between the systems.

5. Summary and conclusions

In this study, wideband parametric speech synthesis with WLP was experimented. WLP showed to produce higher quality wideband synthetic speech compared to conventional LP. Moreover, wideband synthetic speech with WLP was rated higher than narrowband speech. However, bandwidth, model order, or warping did not show statistically significant differences in speaker similarity.

The effect of warping to the quality of synthetic speech was evaluated with two different values of λ. Although the approximation of the Bark scale has provided the best results in speech coding with WLP, a less warped system yielded the best results in the present tests for parametric speech synthesis.

The experiments showed that WLP-based wideband synthetic speech is rated higher in quality than narrowband speech and wideband LP-based speech. However, the quality differences between the systems were rather small, which we think is an effect of the excitation generation method; due to the excluded HNR modification, the highband excitation was partly too periodic, which was audible especially with the female voice. Interestingly, LP-based wideband speech was not statistically different from LP-based narrowband speech. Possible reason for this is that the re-estimation of high-order models is more difficult with conventional LP because it tends to place many of its LSFs into a frequency range with no relevant formant structure.

Present experiments showed no statistically significant differences between the systems in speaker similarity, although previous studies have suggested such improvements in wideband synthesis. This effect might be also explained by the overly periodic highband excitation, decreasing the similarity of the female voice. However, system D with 50th order WLP and $\lambda = 0.35$ yielded the best mean score, possibly indicating better similarity with increased bandwidth and model order, and the use of warping.

Although warping was demonstrated to be successful in wideband speech synthesis, further work for improving the quality is required, for example by including the HNR modification of the excitation or using a wideband pulse library [16] for excitation generation.

6. Acknowledgements

This research is supported by the EC FP7 project Simple4All (287678), Academy of Finland (1128204, 1218259, 121252, 135003 LASTU), MIDE UI-ART, and Tekes.

7. References

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

[2] Yamagishi, J. and King, S., “Simple methods for improving speaker-similarity of HMM-based speech synthesis”, Proc. ICASSP, pp. 4610–4613, 2010.

[3] Stan, A., Yamagishi, J., King, S. and Aylett, M., “The Romanian Speech Synthesis (RSS) corpus: building a high quality HMM-based speech synthesis system using a high sampling rate”, Speech Commun., 53(3):442–450, 2011.

[4] Imai, S., “Cepstral analysis synthesis on the mel frequency scale”, Proc. ICASSP, vol. 8, pp. 93–96, 1983.

[5] 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. on Audio, Speech, and Lang. Proc., 19(1):153–165, 2011.

[6] Suni, A., Raitio, T., Vainio, M. and Alku, P., “The GlottHMM speech synthesis entry for Blizzard Challenge 2010”, The Blizzard Challenge 2010 workshop, 2010. Online: http://festvox.org/blizzard

[7] Stevens, S.S., Volkmann, J. and Newman, E.B., “A scale for the measurement of the psychological magnitude pitch”, J. Acoust. Soc. Am., 8(3):185–190, 1937.

[8] Markel, J.D. and Gray, A.H., Linear Prediction of Speech, second edition, Springer-Verlag, 1980.

[9] Strube, H.W., “Linear prediction on a warped frequency scale”, J. Acoust. Soc. Am., 68(4):1071–1076, 1980.

[10] Härmä, A., Laine, U.K., “A comparison of warped and conventional linear predictive coding”, IEEE Trans. on Speech and Audio Proc., 9(5):579–588, 2001.

[11] Härmä, A., Karjalainen, M., Savioja, L., Välimäki, V., Laine, U.K. and Huopaniemi, J., “Frequency-warped signal processing for audio applications”. J. Audio Eng. Soc., 48(11):1011–1031, 2000.

[12] Härmä, A., “Implementation of frequency-warped recursive filters”, Signal Process., 80:543–548, 2000.

[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”, Sixth ISCA Workshop on Speech Synthesis, pp. 294–299, 2007.

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

[15] Smith, J.O. and Abel, J.S., “The Bark bilinear transform”, Proc. IEEE WASPAA, pp. 202–205, 1995.

[16] Raitio, T., Suni, A., Pulakka, H., Vainio, M. and Alku, P., “Utilizing glottal source pulse library for generating improved excitation signal for HMM-based speech synthesis”, Proc. ICASSP, 2011, pp. 4564–4567.