Wideband Parametric Speech Synthesis Using Warped Linear Prediction

Tuomo Raitio
Paavo Alku

Antti Suni
Martti Vainio

Aalto University

UNIVERSITY OF HELSINKI

Introduction

Statistical parametric speech synthesizers usually utilize a sampling frequency of 16 kHz

Produces pleasant and intelligible speech

However, 16 kHz speech sounds muffled compared to speech with full audio range:

16 kHz  44.1 kHz 

Can we get benefit from higher sampling rates?

Previous Work

Yamagishi and King [1] achieved enhanced feature extraction and improved speaker similarity at higher sampling rates

Stan et al. [2] found that using speech sampled at 32 kHz or more resulted in better speaker similarity compared 16 kHz speech

Neither did observe any improvements in naturalness or intelligibility

In both studies, mel-cepstral type vocoders were used for feature extraction

Mel-Cepstrum vs. Linear Prediction

Statistical parametric speech synthesis mostly utilizes either mel-cepstral or linear prediction (LP) features

Similar results are achieved with both methods

Mel-cepstrum is warped according to the Mel scale while LP uses a linear frequency scale

Model order of LP need to be set carefully depending on the bandwidth

Linear Prediction

The order of LP is selected by the rule:

p = sampling freq. [kHz] + small integer

16 kHz speech: p = 20

44.1 kHz speech: p = 50

However, the number of actual resonances in speech does not increase linearly as a function of sampling frequency

Warped Linear Prediction (WLP)

Warped Linear Prediction (WLP) was proposed by Strube in 1980 [3]

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

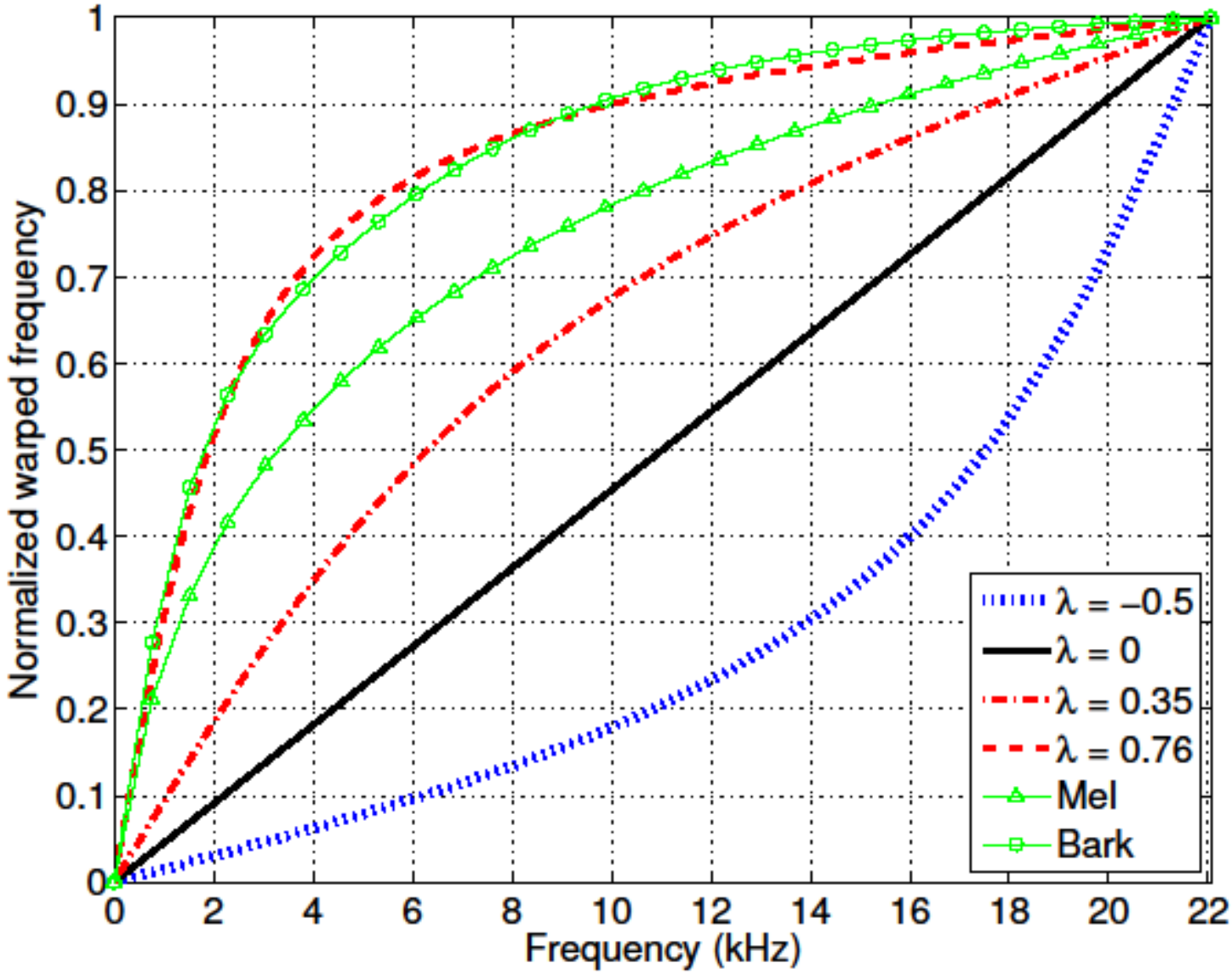
Warped Linear Prediction (WLP) (2)

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

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

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

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


Normalized warped frequency
Frequency (kHz)
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
2
4
6
8
10
12
14
16
18
20
22
λ = –0.5
λ = 0
λ = 0.35
λ = 0.76
Mel
Bark

Using WLP in Speech Synthesis

Benefits of WLP with higher sampling rates:

- Better modeling accuracy in relevant frequencies
- Smaller model order

WLP has not been utilized widely in statistical parametric speech synthesis

Using WLP in Speech Synthesis (2)

We have integrated WLP in our text-to-speech system GlottHMM [4]:

- Separation of vocal tract filter and voice source by glottal inverse filtering
- Modeling of detailed source features
- Using natural glottal flow pulse for generating excitation

Using WLP in Speech Synthesis (3)

Feature	Parameters per frame
Fundamental frequency	1
Energy	1
Harmonic-to-noise ratio	5
Voice source spectrum	10
Vocal tract spectrum	30—50 (LP/WLP)

Experiments

Three different listening tests were conducted in quiet listening booths with high-quality headphones:

1. Quality: Effect of warping in 44.1 kHz speech
2. Quality: Effect of bandwidth, LP model order, and warping for 16 and 44.1 kHz speech
3. Similarity: Effect of bandwidth, LP model order, and warping for 16 and 44.1 kHz speech

A total of 12 native Finnish listeners participated in each test

Experiments – Speech Material and Voices

New Finnish 'Heini' (female) database

- Consisting of 500 phonetically rich sentences and 270 sentences of continuous non-fiction
- A total of 50 642 phone instances

Training of the voices was performed twice: The tree sizes of LSF streams was adjusted in order to roughly match the model complexity between voices

All synthetic speech samples were generated with GlottHMM, either with LP or WLP

Listening Test 1 – Effect of Warping

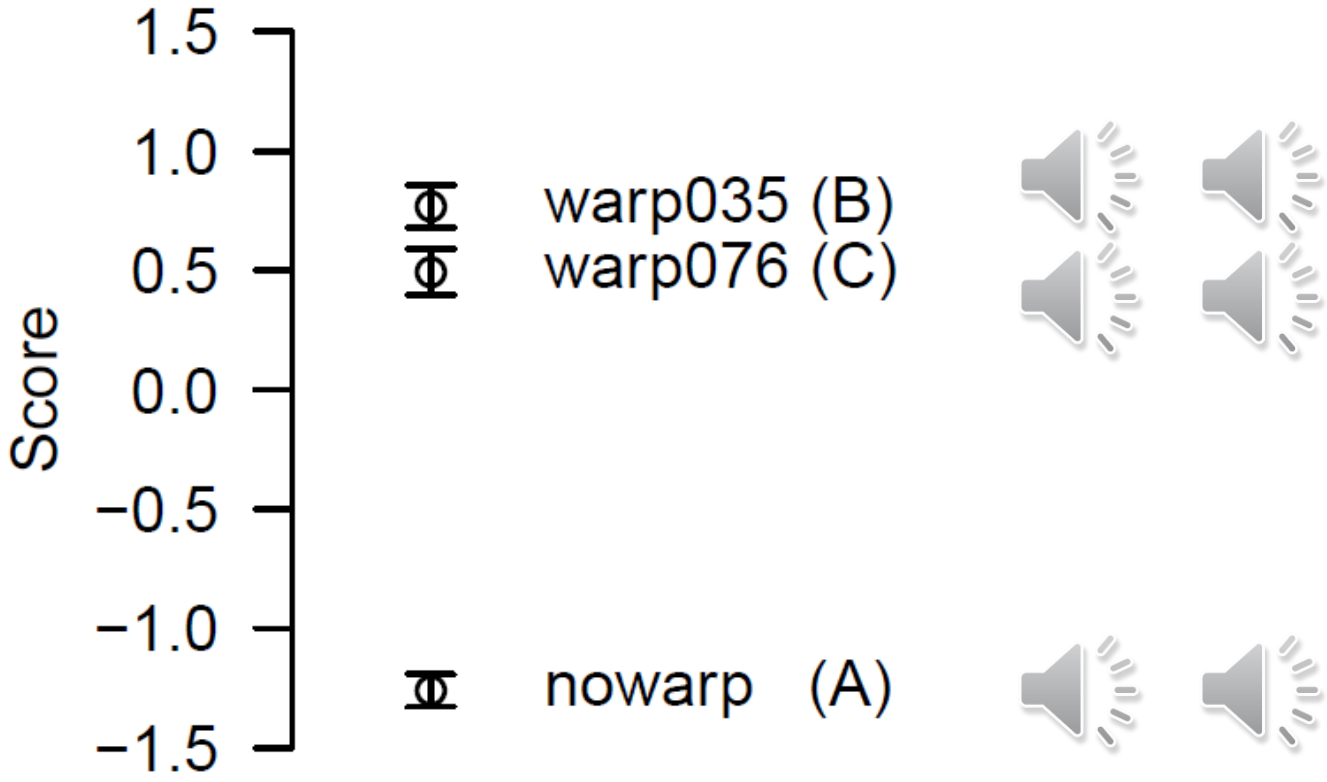
First, the effect of warping on wideband 44.1 kHz speech synthesis was evaluated.

Three systems were included:

A. 30th order LP
B. 30th order WLP ($\alpha = 0.35$)
C. 30th order WLP ($\alpha = 0.76$)

(0.35 gave best results in analysis-synthesis experiments)

(0.76 corresponds approximately to Bark scale)

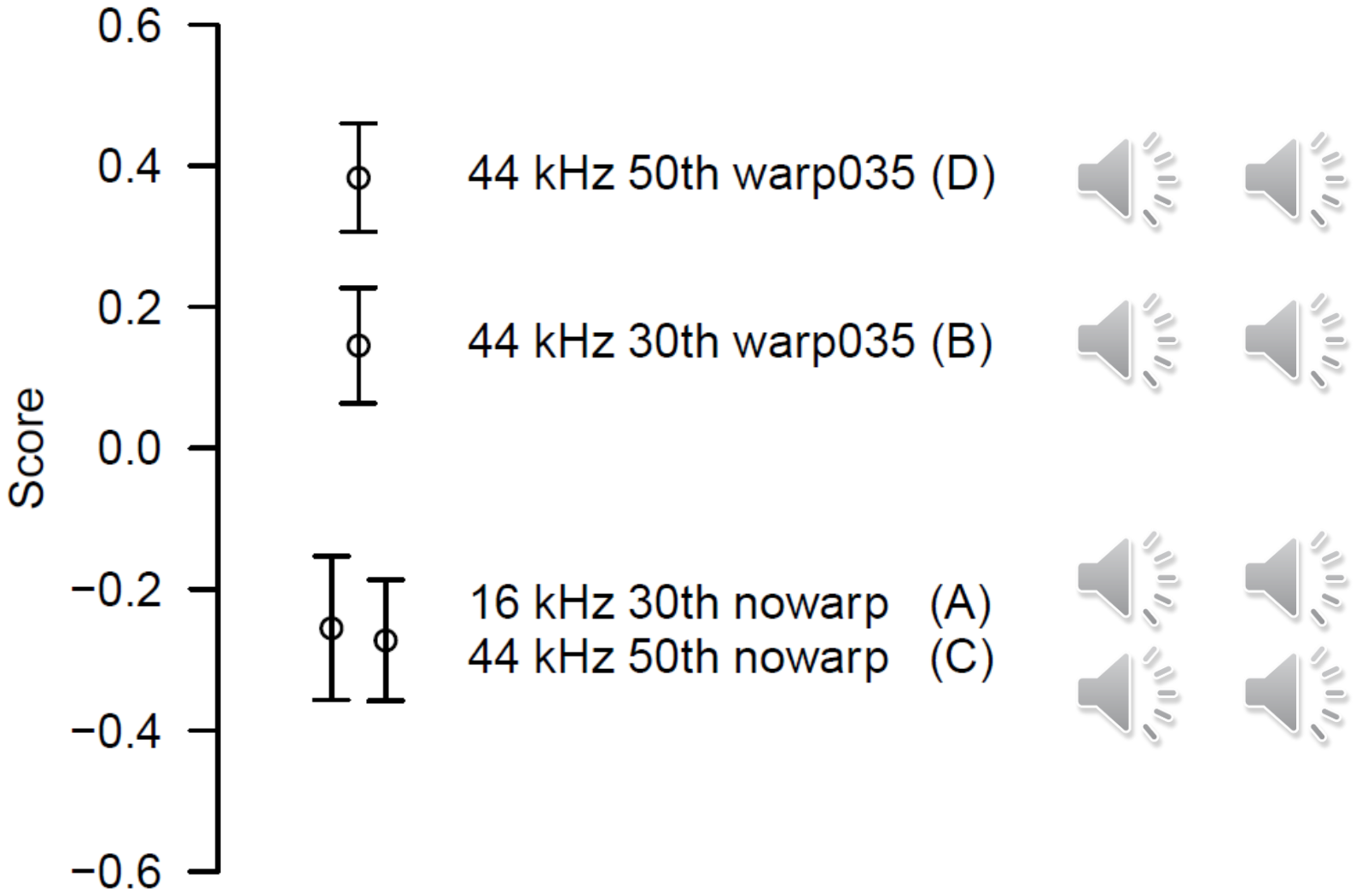

Score
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
warp035 (B)
warp076 (C)
nowarp (A)

Listening Test 2 – Effect of Bandwidth, Model Order and Warping

Four systems were included:

A. 16.0 kHz speech, 30th order LP
B. 44.1 kHz speech, 30th order WLP ($\alpha = 0.35$)
C. 44.1 kHz speech, 50th order LP
D. 44.1 kHz speech, 50th order WLP ($\alpha = 0.35$)

(Voice A: good quality 16 kHz baseline voice)

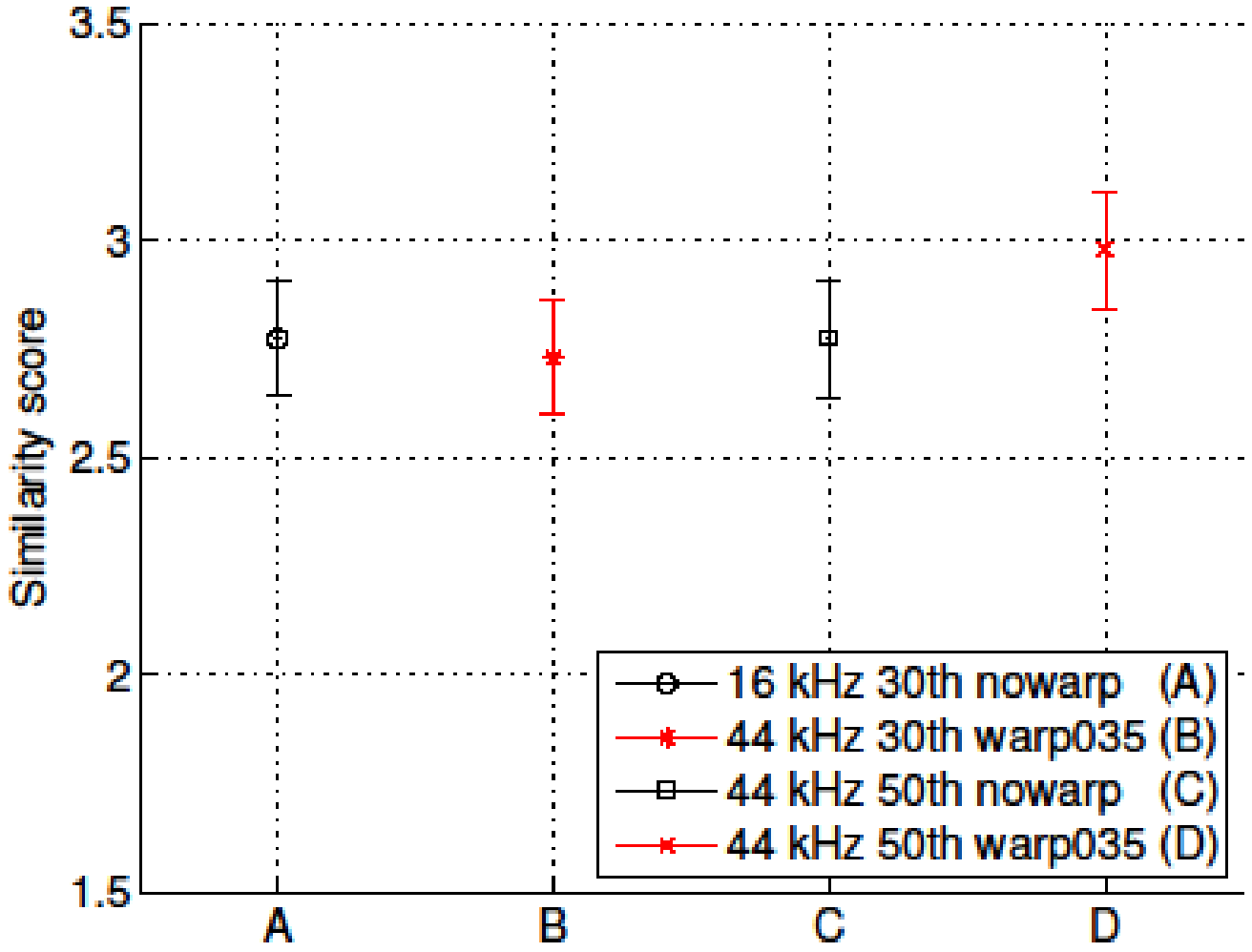

0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
Score
44 kHz 50th warp035 (D)
44 kHz 30th warp035 (B)
16 kHz 30th nowarp (A)
44 kHz 50th nowarp (C)

Listening Test 3 – Speaker Similarity

Again, the same four systems were included:

A. 16.0 kHz speech, 30th order LP
B. 44.1 kHz speech, 30th order WLP (α = 0.35)
C. 44.1 kHz speech, 50th order LP
D. 44.1 kHz speech, 50th order WLP (α = 0.35)

(Voice A: good quality 16 baseline voice)


Similarity score
3.5
3
2.5
2
1.5
A
B
C
D
16 kHz 30th nowarp (A)
44 kHz 30th warp035 (B)
44 kHz 50th nowarp (C)
44 kHz 50th warp035 (D)

Conclusions

1. Wideband synthetic speech with WLP was rated higher than narrowband speech

2. WLP produced higher quality wideband synthetic speech compared to conventional LP

3. Slightly warped systems ($\alpha = 0.35$) yielded the best results in the present tests for parametric speech synthesis

4. Speaker similarity was not statistically different between the voices

References

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

[2] 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.

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

[4] 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.

Thank you for your attention.

Questions?