Using an intelligibility measure to create noise robust cepstral coefficients for HMM-based speech synthesis

Cassia Valentini-Botinhao[1]*, Yan Tang*[2]*, Junichi Yamagishi*[1]*, Simon King*[1]

[1] The Centre for Speech Technology Research, University of Edinburgh, UK
[2]Language and Speech Laboratory, Universidad del País Vasco, Spain
C.Valentini-Botinhao@sms.ed.ac.uk, y.tang@laslab.org,
jyamagis@inf.ed.ac.uk, Simon.King@ed.ac.uk

Abstract

The aim of this work is to increase intelligibility of HMM-based synthetic speech in noisy environments by modifying clean synthetic speech given that noise is known. For that purpose we need a measure for intelligibility of speech in noise that can automatically define the sort of modifications that we need to apply. In previous experiments [1] we have observed that spectrum envelope modifications can have a significant positive impact on the intelligibility of HMM-generated synthetic speech in noise and that the Glimpse proportion measure (GP) [2] is highly correlated with subjective scores under those circumstances.

We have then introduced a method for cepstral coefficient extraction that modifies spectrum envelope based on the GP measure. The GP accounts only for the effect of additive noise, not requiring a reference unmodified speech signal to produce a intelligibility prediction. To control the amount of distortions introduced by the modification we extract cepstral coefficients using an optimization criterion with two terms. The first term accounts for the minimization of the mismatch between natural speech periodogram and magnitude spectrum as modeled by cepstral coefficient, the current criterion used for cepstral coefficient extraction performed at the training stage of the HMM-based speech synthesis framework [3]. The second term accounts for the maximization of an approximated analytical and differentiable version of the GP measure. Using this method we found significant intelligibility gains however not for all tested noise types which indicates that we need a more effective method for controlling distortions [4].

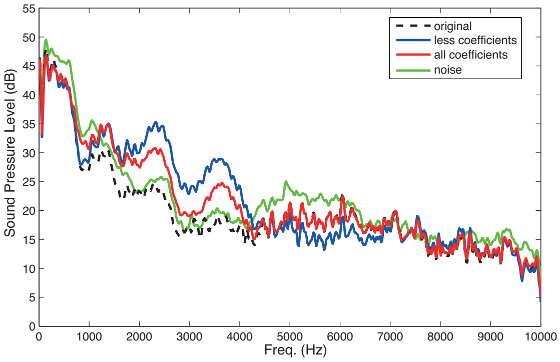

Figure 1: *Long term average spectrum of the original, modified less coefficients (8 coefficients) and modified all coefficients (59) for speech-shaped noise.*

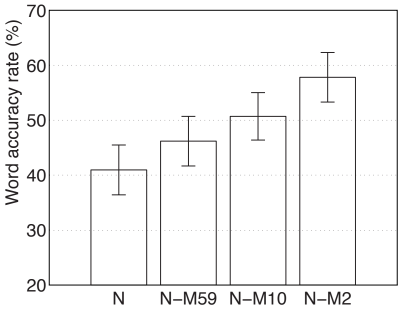

Figure 2: *Word accuracy scores: original (N) and proposed when all (N-M59), 10 (N-M10) and 2 (N-M2) coefficients were modified.*

In this work we propose to limit the frequency resolution of the modifications, and therefore the amount of distortions, by altering only the first few cepstral coefficients, known to be responsible for the coarse frequency resolution of the spectrum. Fig.1 shows the long term average spectrum of original and modified speech, where we can see the effect that limiting the degrees for freedom has on the spectrum envelope. Listening experiments results as shown in Fig.2 indicates that when we modify less coefficients we can improve intelligibility even further.

1. References

[1] C. Valentini-Botinhao, J. Yamagishi, and S. King, "Can objective measures predict the intelligibility of modified HMM-based synthetic speech in noise?" in *Proc. Interspeech*, Florence, Italy, August 2011.

[2] M. Cooke, "A glimpsing model of speech perception in noise," *J. Acoust. Soc. Am.*, vol. 119, no. 3, pp. 1562–1573, 2006.

[3] K. Tokuda, T. Kobayashi, and S. Imai, "Adaptive cepstral analysis of speech," *IEEE Trans. Speech and Audio Processing*, vol. SA-3, no. 6, pp. 481–489, Nov. 1995.

[4] C. Valentini-Botinhao, R. Maia, J. Yamagishi, S. King, and H. Zen, "Cepstral analysis based on the Glimpse proportion measure for improving the intelligibility of HMM-based synthetic speech in noise," in *Proc. ICASSP*, Kyoto, Japan, March 2012.