High quality synthetic speech on a wide vocal effort continuum: Statistical parametric speech synthesis with glottal pulse library

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

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

1. Introduction

Humans adapt their speech according to auditory environment in order to get the message delivered but without using unnecessary effort. Depending on the context, natural speech might vary from whisper to shouting. This vocal effort continuum is an integral part of human communication, but it is typically not utilized in machine-to-human communication. In order to produce contextually appropriate synthetic speech, the auditory environment and context must be taken into account and speech produced at a corresponding point in the vocal effort continuum.

2. Problem formulation

Modeling speech over a wide vocal effort continuum is not easy. In unit selection synthesis, this would require recording of various large databases along the continuum. In statistical parametric synthesis, two or more smaller databases recorded along the continuum can be used to adapt the normal voice. However, the quality of the adapted voices is not always adequate due to insufficient vocoder techniques, statistical averaging and small amount of data [1].

The problem with any speech synthesis system is that there are too little data, resulting in unseen contexts. However, if some contexts are separated into its components, all combination of the components need not to exist in the speech data. This property is utilized in the recently introduced hybrid unit selection/HMM-based system [2].

3. Hybrid unit selection/HMM-based system

A novel hybrid unit selection/HMM-based method [2], called Glottal Pulse Library technique, is based on using glottal inverse filtering for separating speech signal into a glottal source signal and a vocal tract filter. The estimated glottal source signal is segmented to individual glottal source pulses and parameterised into voice source features. Thus, in the synthesis stage, the excitation signal can be reconstructed by selecting the best matching pulses from the library according to the parameters generated by the HMM. The benefit of such a hybrid unit selection/HMM-based system is that the number of units required for natural sounding synthetic speech is very low since the two components, the glottal source and the vocal tract filter, are separated. Thus, only the varying context or modes of the voice source need to be stored into a pulse library, and the variation due to vocal tract filter is modeled by the HMM.

4. Results

Previously, we have shown that the glottal inverse filtering based vocoder [3] can successfully produce natural and very intelligible Lombard speech [4]. In this paper, we show that the glottal pulse library technique can successfully create a continuum from low to high vocal effort by using small pulse libraries along the continuum. We also explore new parameters for describing the glottal source pulses; we use traditional voice quality parameters, such as H1-H2 [5] and NAQ [6], as a target cost in the selection of the library pulses. Moreover, the pulse library method is faster than our previous implementation of the glottal inverse filtering based vocoder [3].

5. Acknowledgements

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

6. References

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

[2] 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”, ICASSP, 2011, pp. 4564–4567.

[3] 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.

[4] Raitio, T., Suni, A., Vainio, M. and Alku, p., “Analysis of HMM-based Lombard speech synthesis”, Interspeech, 2011, pp. 2781–2784.

[5] Titze, I. and Sundberg, J., “Vocal intensity in speakers and singers”, J. of the Acoustical Society of America 91(5):2936–2946, 1992.

[6] Alku, P., Bäckström, T. and Vilkman, E., “Normalized amplitude quotient for parametrization of the glottal flow”, J. of the Acoustical Society of America, 112(2):701–710, 2002.