Aalto University
School of Electrical
Engineering

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

Tuomo Raitio Paavo Alku
Department of Signal Processing and Acoustics
Aalto University, Espoo, Finland

Antti Suni Martti Vainio
Institute of Behavioural Sciences
University of Helsinki, Helsinki, Finland

UNIVERSITY OF HELSINKI

1 Introduction

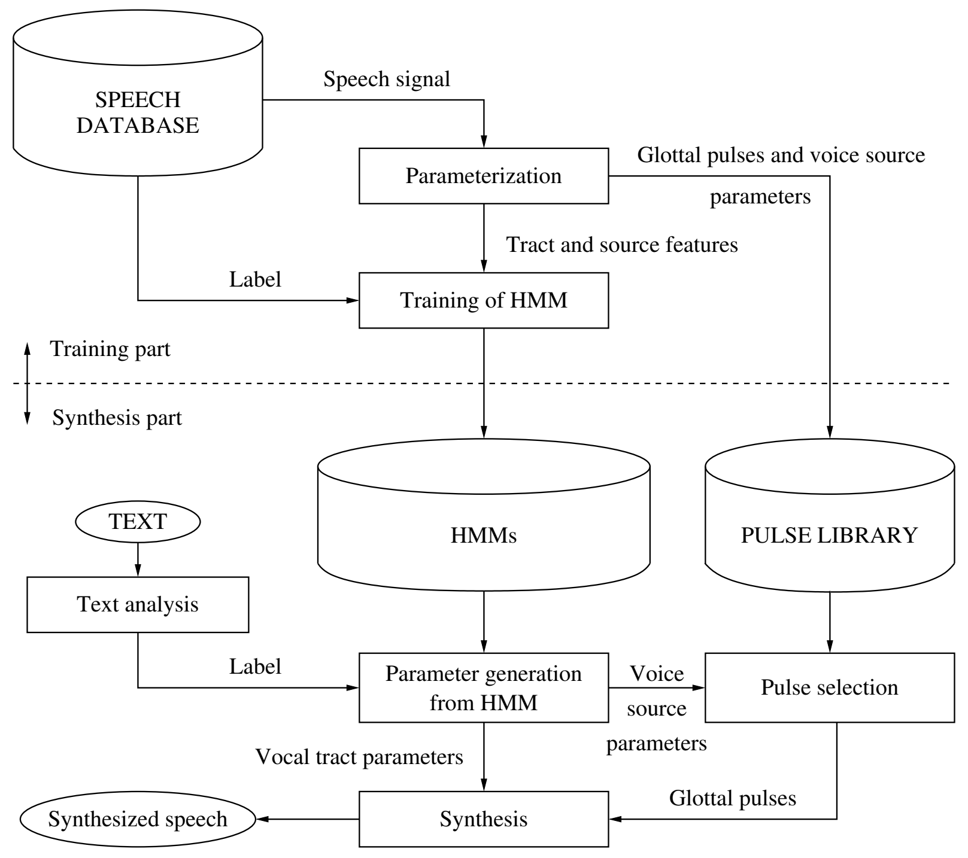
Humans adapt their speech according to the auditory environment in order to get the message delivered without extending unnecessary effort. Depending on the context, natural speech might vary from whisper to shouting. This vocal effort continuum is an integral part of human communication, but it is typically 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.

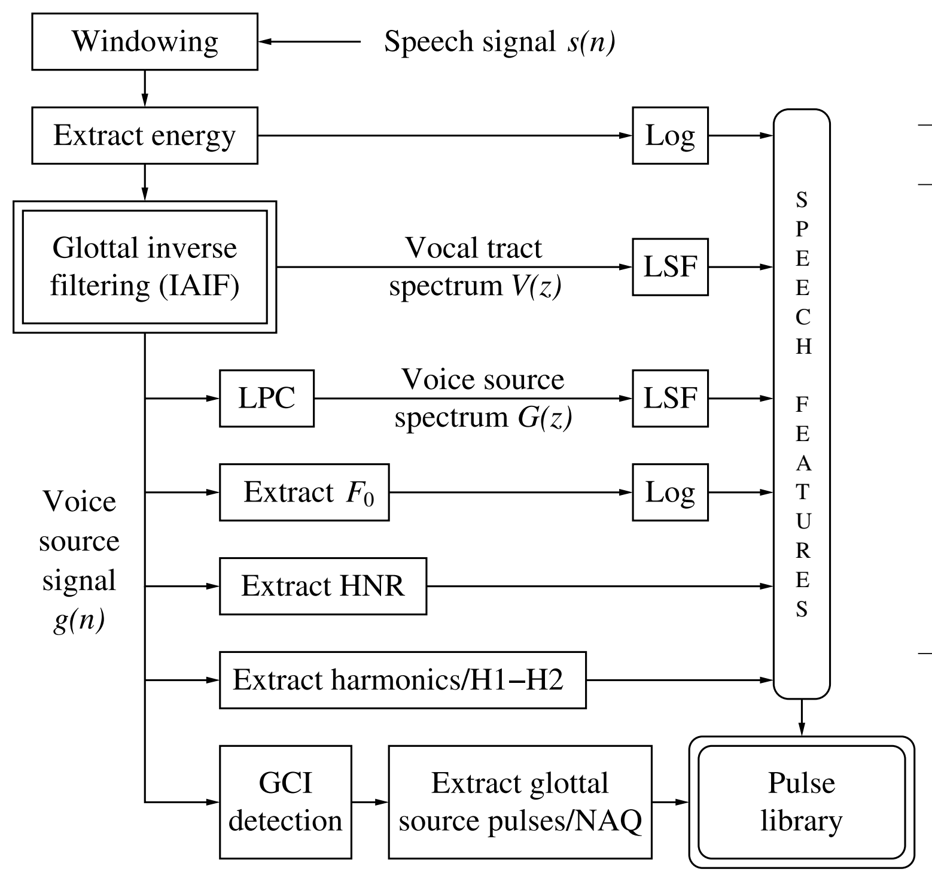
Modeling speech over a wide vocal effort continuum is challenging. In unit selection synthesis, this requires recording of various large databases along the continuum. In statistical parametric synthesis, two smaller databases recorded along the vocal effort continuum can be used to create an adapted voice at an arbitrary point on the continuum by interpolation between the two points or by extrapolating beyond either of the points [1]. However, the quality of adapted voices is not always adequate due to insufficient vocoder techniques and statistical averaging [2]. In addition, the problem with any speech synthesis system is that there are too little data, resulting in unseen contexts.

In this work, we will address the aforementioned issues by utilizing the recently introduced hybrid unit selection/HMM-based system [3].

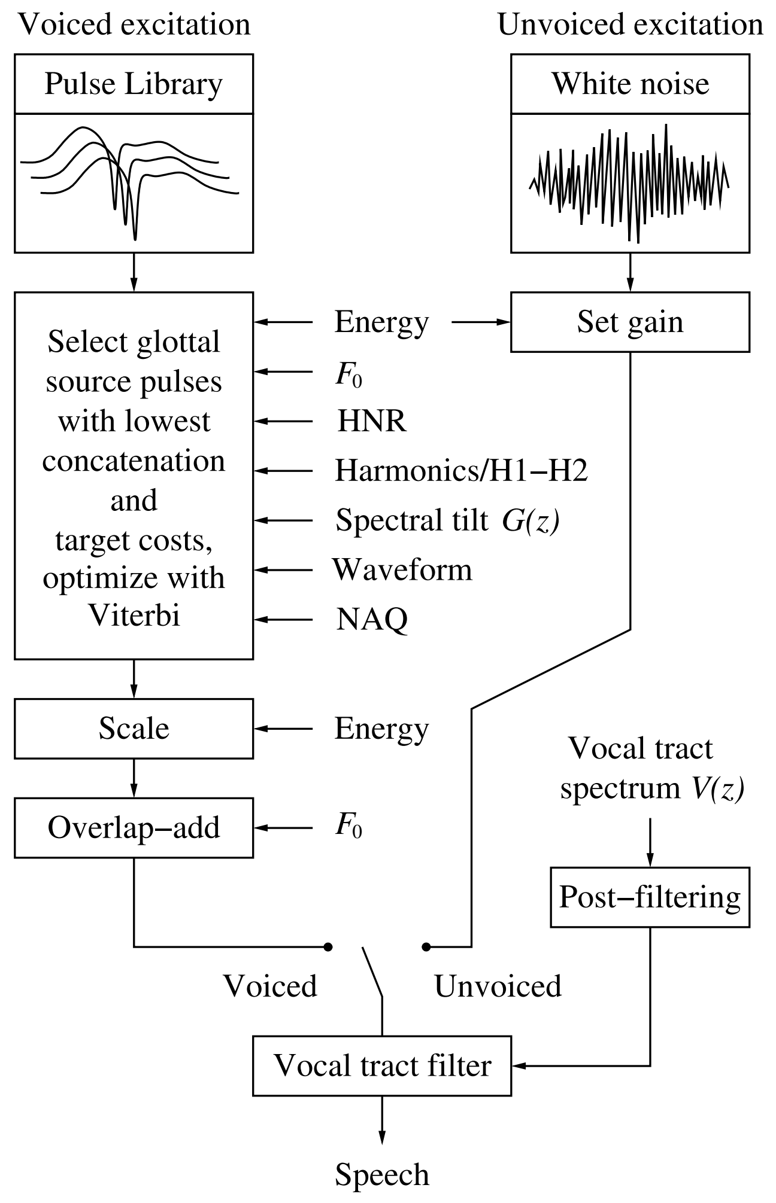
2 Hybrid unit selection/HMM-based system

A novel hybrid unit selection/HMM-based method [3], called Glottal Pulse Library technique, is based on using glottal inverse filtering [4] 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 parameterized 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.





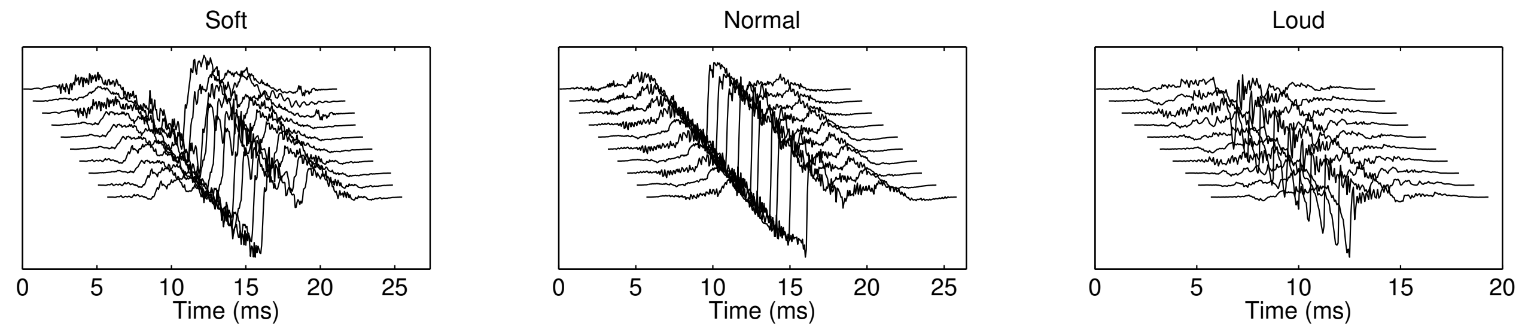
Feature	N° of param.
Fundamental frequency	1
Energy	1
Harmonic-to-noise ratio	5
Harmonic magnitudes	10
Voice source spectrum	20
Vocal tract spectrum	30
NAQ	1
H1–H2	1
+ Pulse library	10–10000 pulses



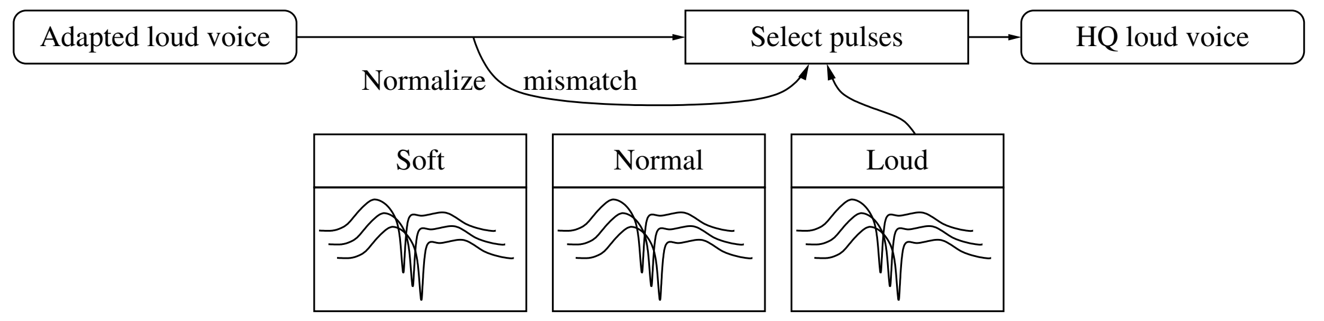
3 Modeling of vocal effort

Previously, we have shown that the glottal inverse filtering based vocoder [5] can successfully produce natural and very intelligible Lombard speech [1]. In this work we demonstrate that the glottal pulse library technique can successfully enhance adapted and interpolated voices on vocal effort continuum, and that conversion of effort can be performed even without HMM methods.

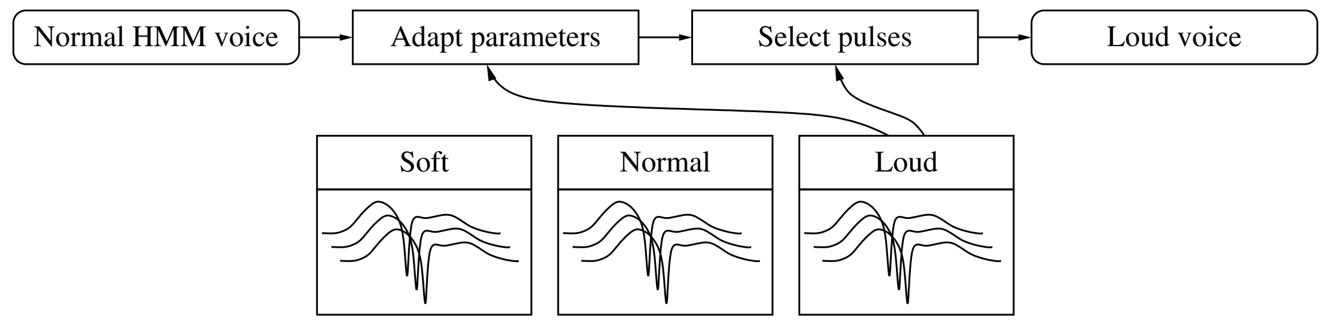
Creating pulse libraries Small glottal source pulse libraries are created from e.g. soft, normal, and loud (or Lombard) speech. For creating a suitable pulse library, only about 5–15 sentences is enough. This will usually lead to a pulse library containing from 2000 to 10 000 pulse segments, depending on the voice and the length of the sentences. However, the number of pulses can be greatly lowered e.g. by using k-means clustering and selecting only the centroids of the clusters. Also, by extracting new, simple, yet relevant voice source features, such as NAQ [6] and H1–H2 [7], the selection of pulses can be made efficient.



Method 1: Adaptation and normalization of pulse library The quality of adapted voice can be enhanced by using a glottal pulse library created from the sentences of matching vocal effort. However, there is always a little mismatch between the synthesis parameters generated from HMMs and the pulse library parameters extracted from natural speech. A solution is to normalize the means of the pulse library parameters according to the synthesis parameters.



Method 2: Normalization of synthesis parameters In order to instantly change the vocal effort of a normal voice without HMM adaptation, the synthesis parameters can be adapted to correspond to the pulse library parameters simply by normalizing the means. Also F0 can be easily transformed to the F0 of the target voice in the logarithmic domain. As a result, a simple unsupervised adaptation method is elaborated, which can produce various voice qualities or even different voices.



4 Conclusions

Glottal pulse library technique can successfully enhance adapted voices on the vocal effort continuum. Moreover, the vocal effort of a normal voice can be changed instantly without HMM adaptation by normalizing the synthesis parameters with the pulse library of a specific voice quality. Speech examples can be found at www.helsinki.fi/speechsciences/synthesis/samples.html or at the Listening Talker workshop 2–3 May 2012 in Edinburgh, Scotland.

References

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

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

[3] 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.

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

[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] 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.

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

Acknowledgements The research leading to these results has received funding from the European Community’s Seventh Framework Programme (FP7/2007–2013) under grant agreement n° 287678, the Academy of Finland, and MIDE UI-ART.

Contact tuomo.raitio@aalto.fi, antti.suni@helsinki.fi, martti.vainio@helsinki.fi, paavo.alku@aalto.fi