COMPARING GLOTTAL-FLOW-EXCITED STATISTICAL PARAMETRIC SPEECH SYNTHESIS METHODS

Aalto University
School of Electrical Engineering

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

Summary:

- **Most potential recent glottal flow signal based excitation methods are studied**
- **Three methods are chosen for comparison: (1) Mean pulse, (2) Mean pulse + 12 principal components, (3) Library of pulses**
- **Listening tests are carried out to determine quality and similarity of synthesis**
- **Mean pulse based methods are better both in quality and similarity**

1 Introduction

HMM-based synthesis [1] is a flexible framework for creating synthetic speech. Despite its several attractive features, HMM-based synthesis is known to suffer from poor voice quality. However, recent advances in vocoding techniques have indicated adequate synthesis quality [1]. One of the key factors for this progress has been the advances in the excitation modeling methods. Several excitation modeling techniques exist, such as impulse train, mixed excitation (e.g. STRAIGHT), etc., that try to model the glottal excitation signal. However, recently there has been an increasing interest in utilizing the natural glottal flow waveform *per se* for synthesis. The goal of this study is to compare the most potential recently proposed glottal-flow-based excitation generation techniques in statistical speech synthesis

2 Speech synthesis system

Parametric speech synthesizer GlottHMM [2, 3] is used

- Based on using IAIF glottal inverse filtering [4]
- Speech is separated into glottal source signal and vocal tract filter
- Source signal is segmented to individual glottal flow pulses

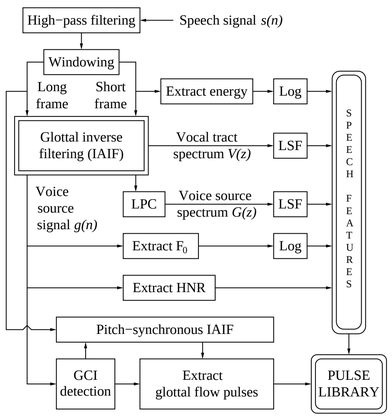
There are several ways to utilize the glottal-flow-pulse library:

- Use only a single natural pulse to create the voiced excitation [2, 5]
- Pulses from the library are selected by minimizing the target cost of the voice source parameters and concatenation cost of the pulse waveforms [3]
- Apply principal component analysis (PCA) to the pulse library and reconstruct pulses by using the principal components (PCs) [6]

Previously, we have shown that both the single pulse technique and the pulse library technique can produce almost natural sounding and very intelligible speech.

In this work we compare these excitation techniques and the PCA-based technique, all built inside the GlottHMM vocoder.

Three methods are chosen for comparison:



Feature	No. of params
Vocal tract sp.	24/30 (f/m)
Energy	1
F0	1
HNR	5
Voice source sp.	5/10 (f/m)
Princ. comp. weights	12
Mean pulse	1 vector
Princ. comp.	12 vectors
Pulse library	~7500 pulses

I. Mean pulse Voiced excitation is constructed using a mean pulse signal, achieved by averaging the pulses in the library (PCA-0). The pulse is interpolated in time and amplitude, noise is added according to HNR in frequency domain, and spectral matching is applied to control the phonation type. Finally, modified pulses are concatenated to create continuous excitation.

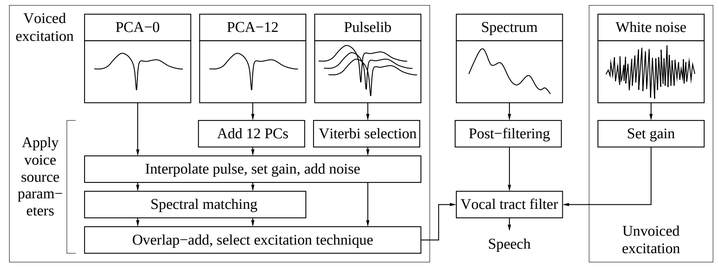
II. Mean pulse + 12 PCs Pulse library is decomposed into the mean of the pulse library and 12 PCs (PCA-12). The principal component weights are trained to HMM system and generated in synthesis. The pulse is constructed using the mean pulse and adding the 12 PCs according to the generated PC weights. Scaling, noise addition and spectral matching is also applied.

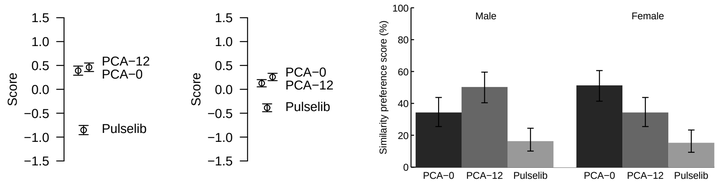
III. Pulse library Individual pulses from the pulse library are selected for each time instant by minimizing the target cost of the voice source parameters and concatenation cost of the pulse waveforms. The selection process is optimized using the Viterbi search. Finally the selected pulses are concatenated to create the excitation signal without spectral matching.

3 Experiments

Quality and similarity were assessed in subjective tests

- Male and female speech
- 20 sentences of the databases were used to build the pulse libraries
- Pulse libraries consisted of 7528 and 7500 pulses
- PCA was applied in order to get the mean pulses, PCs and PC weights
- Quality was measured in comparison category rating (CCR) test
- Similarity was measured in forced choice preference test





4 Conclusions

The results show that the pulse library method is currently not robust enough to yield quality comparable to the single pulse based excitation techniques. The complex unit-selection type optimization of the pulse library technique makes the voice building unpredictable. Although some segments sound very close to original speech, occasional artefacts, reported as "reverberant" or "chorus" type effects in voiced speech, deteriorate the overall quality. This indicates that the smoothness (or regularity) of the resulting speech is of primary importance for the listeners. The study also shows that using PCs in addition to the mean pulse does not increase the quality, corroborating the results in [6]. Although the use of PCs occasionally makes speech more vivid, some artefacts are also produced.

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., 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.

[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., Pulakka, H., Vainio, M. and Alku, P., "HMM-based Finnish text-to-speech system utilizing glottal inverse filtering", in Proc. Interspeech, 2008, pp. 1881–1884.

[6] Drugman, T. and Dutoit T., "The deterministic plus stochastic model of the residual signal and its applications", IEEE Trans. on Audio, Speech, and Lang. Proc., 20(3):968–981, 2012.

Acknowledgements **This research has received funding from the EU's FP7 programme Simple4All under grant agreement n° 287678, Academy of Finland (projects 135003 LASTU programme, 1128204, 1218259, 121252), and MIDE UI-ART.**
Contact **tuomo.raitio@aalto.fi, antti.suni@helsinki.fi, martti.vainio@helsinki.fi, paavo.alku@aalto.fi**