Advances in Voice Modeling for Speech Restoration

O. Schleusing, M. Bertschi, P. Renevey

The goal of this project is the restoration of pathological speech for patients with severely disordered voice. A novel approach for glottal inverse filtering has been developed to allow the separation of recorded speech into pathological glottal components to be discarded and articulatory formants to be retained for restoration. This is achieved by joint optimization of glottal and articulatory models, whereby the undesired pathological voice characteristics are captured in appropriate voice model parameters and the distortion of the articulation parameters is reduced.

Due to its non-invasive nature, glottal inverse filtering (GIF) has recently gained increased popularity among researchers and has become an important speech processing topic in areas such as fundamental speech research, medical applications or general speech technology applications (e.g. speech codecs, speech synthesis). In the context of CSEM research on voice restoration, a novel method for GIF has been developed that is particularly robust to voice characteristics found in pathological voice such as aspiration noise, jitter and shimmer.

Most GIF methods such as linear prediction (LP) use a simple, time-invariant model of the glottal source. Thus, a significant amount of the highly-volatile glottal excitation is captured in the estimated articulatory formant parameters leading to a bias and increased variance of the estimator (see Figure 2). However, GIF uses a more comprehensive model of speech production by introducing a higher dimensional and time-variant model of the glottal contribution. By jointly applying physiological constraints on the formant trajectories, the bias and variance of the estimated formants is reduced, resulting in an increased quality of the resulting restored speech signal.

The estimation of the model parameters is carried out pitch-synchronously using a global optimization technique named differential evolution. For each glottal cycle, a population of glottal and formant parameter sets converges to an optimal set by minimizing the error between the recorded signal and a synthetic signal generated from the respective set of parameters. A general problem in GIF is the overlapping of the resonances of consecutive glottal cycles (see Figure 1). To mitigate this effect, the optimization is performed in an iterative manner. The outcome of an initial optimization of the cycle at time *k* is used to compensate the effect of its overlapping in the initial optimization of the next cycle at time *k+1*. Subsequently, the effect of the cycle at time *k+1* is annihilated during a second, fine-grained optimization of the cycle at time *k*. This sequence is continued for each glottal cycle by using the estimates of the respective previous cycles to obtain optimal solutions.

In a further improvement, parameter constraints are introduced that benefit from physiological constraints of speech production. In particular, it is well known that the parameterization of the articulation follows a significantly slower evolution over time. Thus, it becomes possible to predict future formant frequency values and to incur a weighting factor in the cost function to penalize formant frequencies deviating significantly from the predicted trajectory.

In the currently developed system of speech restoration, a healthy-like speech signal is reconstructed by applying the articulation filter parameters estimated from a pathological speaker to a synthetically generated glottal excitation. This new glottal excitation signal is constructed by concatenating glottal cycles chosen randomly from a database [1], which had previously been built by inverse filtering of healthy speech using the same optimization method described above.

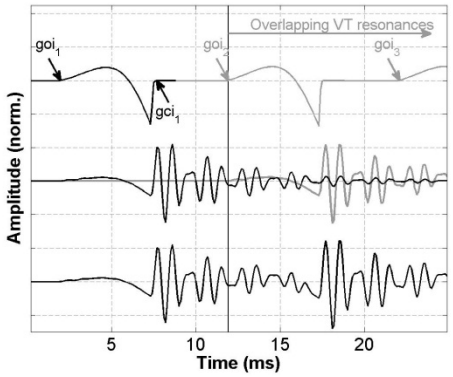

Figure 1: Overlapping resonances of consecutive glottal cycles. The proposed iterative compensation scheme effectively increases the size of the analysis frame.

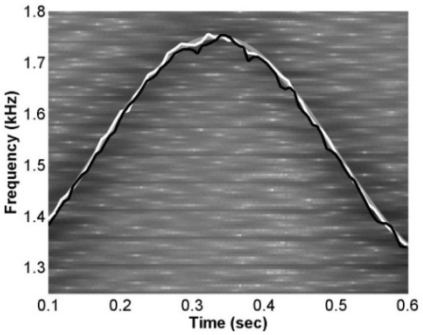

Figure 2: Section of a spectrogram of a synthetically generated vowel transition, overlaid with the centre frequency of formant F2 (thick grey line) and its estimates using the proposed method (white line) and standard linear prediction (black line). The proposed method exhibits a reduced bias and variance. Observe especially the high dependency of the LP method on the underlying harmonic signal structure, which is significantly reduced by the dedicated glottal model of the proposed method.

This work is being currently prepared for publication in an international journal. CSEM thanks Onco-Suisse and the Gebert-Rüf-Stiftung for funding, and Dr. J.-M. Vesin of EPFL for supervising the on-going PhD thesis on this topic.

[1] O. Schleusing, R. Vetter, P. Renevey, J.-M. Vesin, V. Schweizer, "Prosodic speech restoration device: glottal excitation restoration using a multi-resolution approach" Communications in Computer and Information Science: Biomedical Engineering Systems and Technologies, Springer (2011) 177–188