Controlling Voice Source Parameters to Transform Characteristics of Synthetic Voices

João P. Cabral, Julie Carson-Berndsen

School of Computer Science and Informatics, University College Dublin, Ireland
joao.cabral@ucd.ie, julie.berndsen@ucd.ie

1. Introduction
The success of the speech communication process depends not only on the intelligibility of the speech transmitted to the listener but also on how the message is spoken. An important aspect that carries underlying information in speech besides the linguistic content is the type of voice. Humans change instinctively or intentionally their voice depending on their mood, the environment, the listener, the feelings they want to transmit, etc. In order to take advantage of this valuable aspect of speech in applications of synthetic voices for spoken communication it is necessary that the computer can produce a high variability of voices and that it can predict an appropriate voice based on environmental cues, including feedback information about the listener. This work fits into the subject of modelling and transforming acoustic aspects of speech for controlling the type of synthetic voice. The goal is to accurately model an important acoustic component of speech related to voice characteristics which is aspiration noise. This noise signal results from the turbulence of air passing through the glottis during human speech production. It can be represented by an amplitude modulated Gaussian noise, which depends on the glottal volume velocity and glottal area. For example, this modulation effect is more important in breathy voice than modal because the vocal folds usually do not completely close for breathy unlike modal.

Index Terms: Voice transformation, aspiration noise, breathy

2. Voice Transformation Method

2.1. Glottal Spectral Separation for Analysis-Synthesis
In this work, the analysis-synthesis method called Glottal Spectral Separation (GSS) [1] is used to transform a modal voice into breathy. This technique was chosen because it permits to control parameters of an acoustic glottal source model, the Liljencrants-Fant (LF) model [2], and performed well in voice transformations. However, the LF-model does not represent the noise characteristics of the voice source, in particular the aspiration noise. For this reason, we combined the GSS method with a technique for modelling this type of noise.

2.2. Aspiration Noise Modelling
The aspiration noise was estimated using an harmonic-stochastic model of speech (HNM) The harmonic and stochastic components were firstly separated from the speech signal using the UPC tools (http://www.talp.upc.edu/talp/index.php/resources/tools/). Then, the harmonic signal was subtracted from the speech signal to obtain the noise component. Finally, the aspiration noise was estimated from the noise by LPC inverse filtering.

After estimating the aspiration noise it is necessary to model its amplitude modulation effect. The modulation function is calculated by using the Hilbert transform method of envelope detection. In this work, this envelope is initially parameterised using a polynomial fitting technique. Then, a triangular function is obtained from the polynomial representation in order to obtain a more robust and accurate estimate of the noise envelope. A great advantage of the triangular representation compared to typical functions that have been used to represent the envelope, such as the symmetric Gaussian and Hanning windows, is that it better represents different shapes (including asymmetric shapes) of the energy envelope. This flexibility is similar to that of using a glottal source signal representation of the energy envelope [3], but the first is simpler because does not require estimation of the glottal signal. For example, by using this function is possible to adjust the envelope shape depending on the transformations of the glottal parameters.

3. Experimental Results

The values of the voice quality parameters of the LF-model (open quotient, speed quotient and return quotient) calculated for an utterance (modal voice) were modified to obtain parameter contours with mean values equal to those measured on the target voice, similarly to [1]. We additionally modified the mean values of the $F_0$ and $HNR$ contours to improve the voice conversion. $HNR$ was estimated as the energy ratio between the harmonic and noise components of the HNM.

Results of a perceptual experiment showed that the GSS method combined with the aspiration noise model significantly improved the naturalness of synthetic speech and transformation of modal voice into breathy, compared with the baseline GSS method which only used the LF-model to represent the excitation signal.

4. Acknowledgements
This research is supported by the Science Foundation Ireland (Grant 07/CE/I1142) as part of the Centre for Next Generation Localisation (www.cngl.ie) at University College Dublin.

5. References