Vowel Enhancement in Early Stage Spanish Esophageal Speech Using Natural Glottal Flow Pulse and Vocal Tract Frequency Warping

Rizwan Ishaq[1], *Dhananjaya Gowda*[2], *Paavo Alku*[2], *Begoña García Zapirain*[1]

[1]Deustotech-LIFE, University of Deusto, Bilbao, Spain
[2]Aalto University, Dept. of Signal Processing and Acoustics, Finland

rizwanishaq@deusto.es, dhananjaya.gowda@aalto.fi, paavo.alku@aalto.fi, mbgarciazapi@deusto.es

Abstract

This paper presents an enhancement system for early stage Spanish Esophageal Speech (ES) vowels. The system decomposes the input ES into neoglottal waveform and vocal tract filter components using Iterative Adaptive Inverse Filtering (IAIF). The neoglottal waveform is further decomposed into fundamental frequency F_0, Harmonic to Noise Ratio (HNR), and neoglottal source spectrum. The enhanced neoglottal source signal is constructed using a natural glottal flow pulse computed from real speech. The F_0 and HNR are replaced with natural speech F_0 and HNR. The vocal tract formant frequencies (spectral peaks) and bandwidths are smoothed, the formants are shifted downward using second order frequency warping polynomial and the bandwidth is increased to make it close to the natural speech. The system is evaluated using subjective listening tests on the Spanish ES vowels /a/, /e/, /i/, /o/, /u/. The Mean Opinion Score (MOS) shows significant improvement in the overall quality (naturalness and intelligibility) of the vowels.

Index Terms: speech enhancement, glottal flow, analysis synthesis vocal tract, spectral sharpening, warping

1. Introduction

The removal of the larynx after a Total Laryngectomy (TL), changes the speech production mechanism. The trachea which connects the larynx and lungs for air source is now connected to a stoma (hole on neck) for breathing. The vocal folds which resided in larynx are no more available. After TL, there is no voicing and air source for speech production. Therefore alternative voicing and air source are needed for speech restoration. Three methods are available for this purpose, i) Esophageal Speech (ES), ii) Tracheo-Esophageal Speech (TES), and iii) Electrolarynx (EL). ES and TES both use a common voicing source, the Phyarngo-Esophageal (PE) segment, but with a different air source, while EL uses external devices for voicing source with no air source. The ES is preferred over other methods, because it does not require surgery (TES) or external devices (EL). ES involves, however, a low pressure air source, and an irregular PE segment vibration which results in low quality and low intelligible speech. Compared to the production of normal speech according to the source-filter model [1], the voicing source in ES is severely altered and does not have any fundamental frequency or harmonic components. The vocal tract filter is also shortened in ES. The ES can be enhanced by transforming the source and filter components to those of normal speech using signal processing algorithms.

In previous studies ES is typically decomposed into its source and filter components using Linear Predication (LP) based analysis-synthesis techniques. Based on this assumption the authors in [2, 3] replaced the voicing source with the Liljencrants- Fant (LF) voicing source, and reported significant enhancements. Fundamental frequency smoothing and correction with the synthetic LF source model were used for quality enhancement also in [4]. ES enhancement based on formant synthesis has also shown significant improvement in intelligibility [5, 6]. In [7] the source and filter components were modified by replacing the source with the LF model and increasing the bandwidth of filter formants for better quality speech. Statistical conversion from ES to normal speech has also improved intelligibility, but requires more ES data [8]. Some other not so common approaches are based on Kalman filtering [9, 10, 11, 12], and modulation filtering enhancement [13, 14].

Almost all methods available in the literature assume that the fundamental frequency of ES can be estimated accurately. The voicing source signal is then modified with the synthetic LF model voicing source. The vocal tract formants are typically considered to be the same as in normal speech signals. In reality, however, the fundamental frequency of ES is highly irregular and the voicing source resembles whispered speech. Moreover, formants center frequencies are affected by the shortening of vocal tract length due to surgery. In order to deal with these deficiencies, this paper proposes an ES enhancement method based on the GlottHMM single pulse synthesis [15, 16, 17]. The system decomposes ES into neoglottal waveform and vocal tract filter components using Iterative Adaptive Inverse Filtering (IAIF) [18]. Natural glottal pulse extracted from real speech is used to construct the glottal waveform by borrowing F_0 curve and HNR from normal speech. The vocal tract filter is also modified by smoothing the spectral peaks and their bandwidths. The spectral peaks of the vocal tract filter are also moved to lower frequencies in order to compensate the rising of formant in ES. The formant bandwidths are also increased for better quality speech. The system is validated with Spanish Esophageal Vowels subjectively using the Mean Opinion Score (MOS). The paper in next section describes the system in detail. The subsequent sections contain results, discussion and finally conclusions.

2. System Description

The proposed system, shown in Figure 1, is divided into three main components, i) analysis, ii) transformation, and iii) synthesis. The analysis part decomposes the voiced speech frame into its source and filter components. The transformation provides the modified source and filter components. Finally the modified components are combined in the synthesis part to generate enhanced ES.

SLPAT 2015, 6th Workshop on Speech and Language Processing for Assistive Technologies, pages 55–59,
Dresden, Germany, 11 September, 2015.

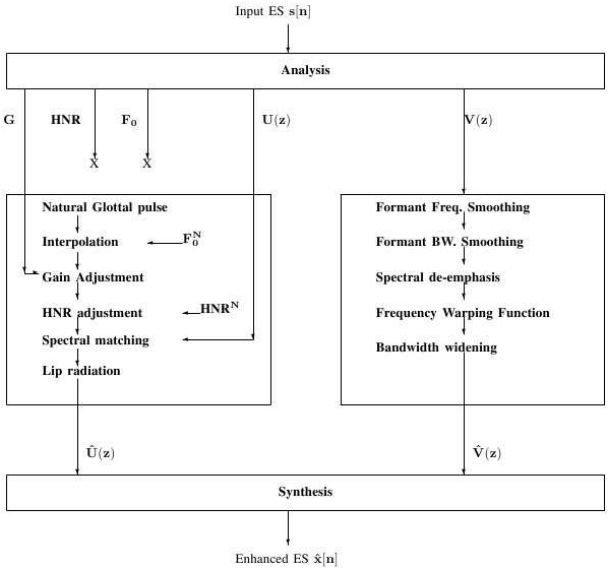

Figure 1: *Proposed enhancement system.*

2.1. GlottHMM based analysis

The goal of the analysis part of the system is to decompose the ES signal into a neoglottal source signal and a vocal tract spectrum. The input speech signal $s[n]$ is first passed through highpass filter $h_{hp}[n]$ with a cutoff frequency of 70 Hz.

$$s_h[n] = s[n] * h_{hp}[n] \tag{1}$$

where $s_h[n]$ and $*$ are the highpass filtered speech signal and a convolution operator, respectively. The highpass filtered signal $s_h[n]$ is then windowed using a rectangular window of size 45-ms, with 5-ms frame shift.

$$x[n] = s_h[n]w[n] \tag{2}$$

where $w[n]$ is the rectangular window. Firstly the log energy G of frame is extracted using,

$$G = log(\sum_{n=0}^{N-1} x^2[n]) \tag{3}$$

where N is the number of samples in the frame. Glottal Inverse Filtering (GIF) is then used to separate the frame into a neoglottal source signal and a vocal tract spectrum. The automatic inverse filtering, IAIF is used [18]. IAIF estimates vocal tract and lip radiation using all-pole modeling and then iteratively cancel these components. In simplified form, the neoglottal source signal:

$$U(z) = \frac{X(z)}{V(z)R(z)} \tag{4}$$

where $U(z)$, $X(z)$, $V(z)$ and $R(z)$ are the z-transforms of neoglottal source signal $u[n]$, speech signal $x[n]$, vocal tract impulse response $v[n]$, and lip radiation response $r[n]$ respectively. The estimated neoglottal source signal $u[n]$ is parametrized into fundamental frequency F_0, Harmonic to Noise Ratio (HNR) and neoglottal source spectrum $U(z)$. The autocorrelation of the neoglottal source signal $u[n]$ is used for F_0 estimation. The HNR is estimated using the upper and lower

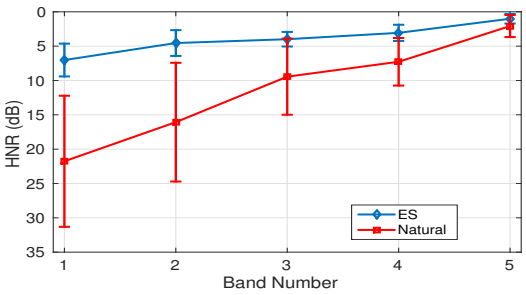

Figure 2: *HNR of ES and natural speech.*

smoothed spectral envelopes ratio to determine the voicing degree in the neoglottal voicing source signal $u[n]$ for five frequency bands [15]. In short the analysis part of the system provides for each frame the following, i) Frame energy G, ii) vocal tract spectrum $V(z)$ (LP order 30), iii) F_0, iv) HNR and v) neoglottal source spectrum $U(z)$ (LP order 10).

2.2. ES to normal speech transformation

The parameters obtained from the analysis are transformed into natural speech parameters. The neoglottal signal and vocal tract are modified independently.

2.2.1. Neoglottal source signal enhancement

The neoglottal source signal $u[n]$ is the most effected speech component in ES. Therefore the parameters of this signal are replaced with any arbitrary natural speech signal for a better glottal source signal. The natural glottal pulse which is extracted from normal speech is first interpolated using the cubic spline interpolation by replacing the frame original F_0 with natural speech F_0^N . The interpolated glottal pulse voicing source is then multiplied with the smooth gain G and the natural speech HNR is then used to add noise in the frequency domain for naturalness according to the following steps:

- Taking FFT of the neoglottal waveform,
- Adding random components (white Gaussian noise) to real and imaginary part of FFT according to HNR,
- Taking IFFT of noise added neoglottal waveform

$$U_{syn}(z) = 10^G G(z) + Q(z) \tag{5}$$

where $U_{syn}(z)$ is the synthetic glottal source, $G(z)$ is the natural glottal pulses source, and $Q(z)$ is HNR based noise component. Figure 2 shows the mean value of HNR for all voiced frames along with standard deviation. The figure indicates that HNR of ES is greatly different from that of normal speech. Therefore, it is justified to replace the HNR of ES with the HNR of normal speech in the vowel enhancement system. In order to adjust the spectrum of neoglottal waveform to the spectrum of the target waveform, the former is filtered with following IIR filter:

$$H_m(z) = \frac{U(z)}{U_{syn}(z)} \tag{6}$$

where $U(z)$ and $U_{syn}(z)$ are the LP spectra of the original and synthetic neoglottal waveform, respectively. The lip radiation is applied to the spectrally matched neoglottal waveform $\hat{u}[n]$:

$$\hat{u}[n] = \hat{u}[n] - \alpha\hat{u}[n-1], \quad 0.96 < \alpha < 1 \tag{7}$$

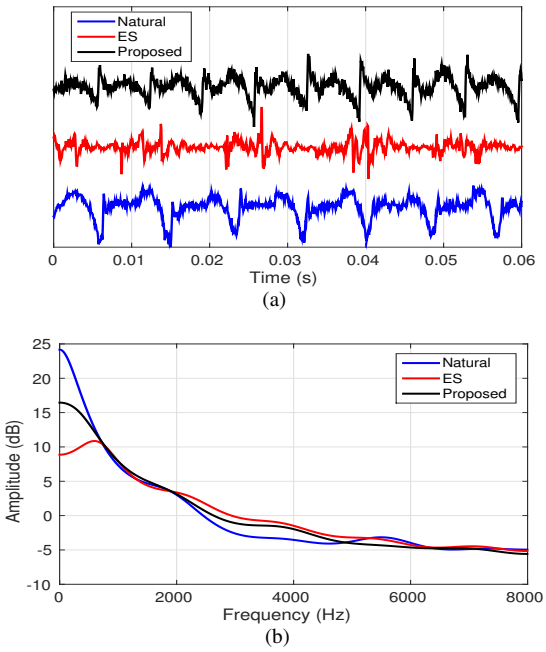

Figure 3: *Glottal excitations (computed from the vowel /a/) in the time domain (a) and in the frequency domain (b).*

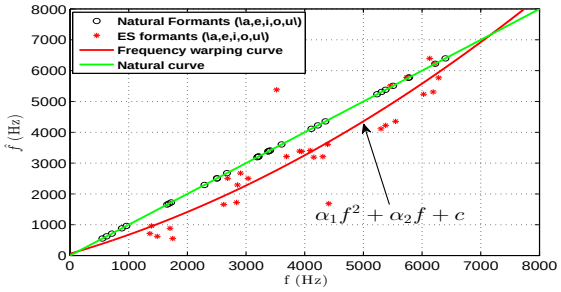

Figure 4: *Frequency Warping Function (FWF) curve.*

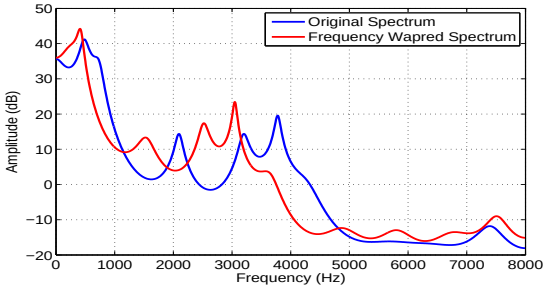

Figure 5: *Frequency warped spectra.*

where $\hat{u}[n](\hat{U}(z))$ and $\alpha(0.98)$ are the modified neoglottal waveform and lip radiation constant, respectively.

Figure 3(a) shows time-domain examples of glottal excitations of natural speech and ES together with a waveform computed with the proposed enhancement system. It can be seen that the proposed system is capable of producing a glottal excitation that is highly similar to that of natural speech. As shown in Figure 3(b), the spectral slope of the excitation waveform generated by the proposed method is also close to that of natural speech, especially at low frequencies, but the generated spectrum also retains the spectral slope of ES at higher frequencies.

2.2.2. *Vocal tract modification by nonlinear frequency warping*

The vocal tract spectrum of ES has the following characteristics, i) higher frequencies are emphasized more compared to lower frequencies, ii) spectral resonances (formants) are moved to higher frequencies, and iii) resonance bandwidths are reduced in comparison to normal speech vowels. To cope with the higher frequency emphasis, a de-emphasis filter is applied to the vocal tract spectrum. The resulting vocal tract transfer function is then expressed as:

$$H_{enh}(z) = \frac{1+\alpha z^{-1}}{1+\sum_{p=1}^{P} a_p z^{-p}}, \quad 0.95 < \alpha < 1 \qquad (8)$$

where P is the order of the all-pole vocal tract filter and α is the de-emphsis constant.

Because formants of ES are moved upward in frequency, a procedure is needed to adjust them to coincide more closely with the formant values of normal speech. For such a procedure, we used a second order Frequency Warping Function (FWF) $\zeta(f)$ defined as:

$$\zeta(f) = \alpha_1 f^2 + \alpha_2 f + c \qquad (9)$$

where $\alpha_1 = 6.079 \times 10^{-5}$, $\alpha_2 = 0.5553$, and $c = 60.280$.

$$\hat{f} = \beta\zeta(f), \quad \beta = 1, f = 0 \rightarrow \frac{f_s}{2} \qquad (10)$$

where $\hat{f}$ and f, are warped and original frequencies, and β is a constant. Figure 4 demonstrates FWF using first four formants of vowels (/a/, /e/, /i/, /o/, /u/) extracted from normal speech (x-axis) and ES (y-axis). The obtained frequency warping, applicable for a general formant mapping between normal speech and ES, is shown in Figure 5. In order to expand the formant bandwidths, exponential windowing is used for the vocal tract filter coefficients as follows [19]:

$$H_s(z) = \frac{1+\sum_{p=1}^{P} \gamma^p a_p z^{-p}}{1+\sum_{p=1}^{P} \eta^p a_p z^{-p}}, \quad 0.90 < \gamma, \eta < 1 \qquad (11)$$

where γ and η are constants controlling the spectral bandwidth.

If $\gamma > \eta$ bandwidth of formants increase, otherwise it decreases (i.e. formants are sharpened). For the purpose of the present study, $\eta(0.97)$ is always smaller than $\gamma(0.99)$ in order to increase formant bandwidths.

2.3. Synthesis of enhanced speech

The synthesis part involves convolving the modified neoglottal waveform and the impulse response of the vocal tract filter yielding the enhanced version of ES $\hat{x}[n]$;

$$\hat{x}[n] = \hat{v}[n] * \hat{u}[n] \qquad (12)$$

where $\hat{u}[n]$ and $\hat{v}[n]$ are the modified neoglottal waveform and vocal tract impulse response, respectively.

3. System Evaluation

The system was evaluated with ES vowels of Spanish (/a/, /e/, /i/, /o/, /u/) recorded in speech rehabilitation center. The data

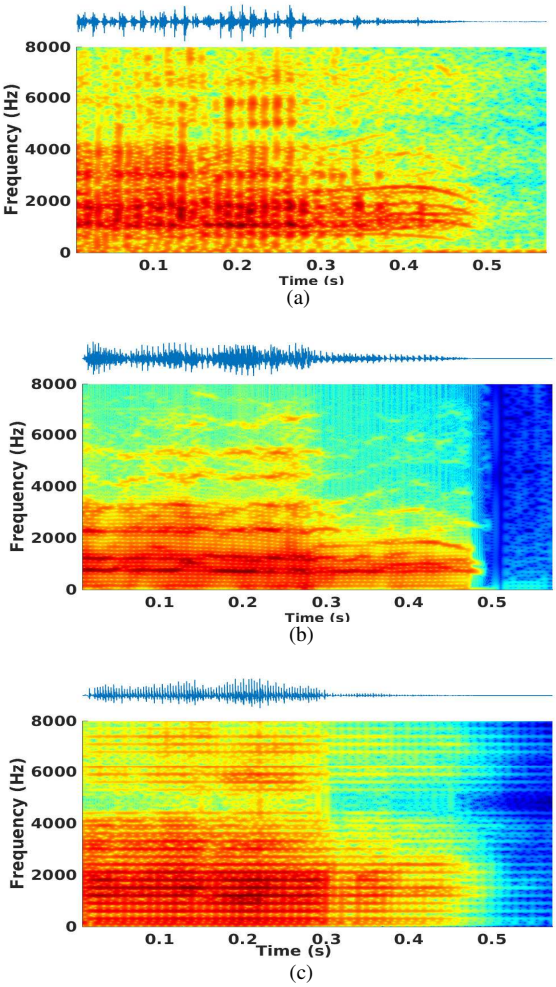

Figure 6: *Spectrograms of the vowel /a/ for different processing types: unprocessed (a), processed with the proposed system (b), processed with the reference system (c) [7]*

was collected from five early stage male ES talkers by asking them to utter each vowel four times. Due to lack of female patients in the rehabilitation center, only male speakers were involved in the study. The speech sounds were sampled with 44.1 kHz from which the data was down-sampled to 16 kHz for computational efficiency.

The system performance is visually demonstrated with spectrograms in Figure 6. In this figure, and also later in Figures 7 and 8, the proposed system is compared with a reference system based on using the LF source and formant modification with a bandwidth extension system [7]. It can be seen from Figure 6 that the spectrogram computed from the enhanced vowels by the proposed system shows a clearer formant and harmonics structure in comparison to ES and the reference system.

3.1. Subjective listening evaluation

Two subjective listening tests were conducted. The first one was a quality evaluation based on the Mean Opinion Score (MOS) which is a widely used perceptual quality test of speech based on a scale from 1 (worst) to 5 (best). In this test, the listeners heard original ES vowels and the corresponding enhanced ones,

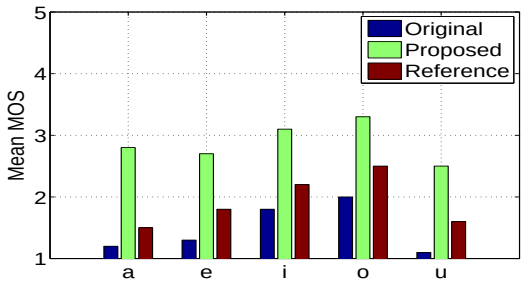

Figure 7: *Results of the MOS test for all the vowels.*

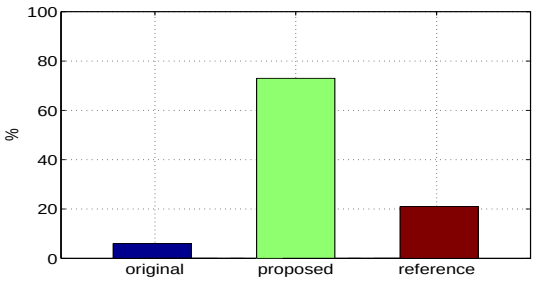

Figure 8: *Results of the preference test.*

processed by both the proposed and the reference method, in a random order and they were asked to grade the quality of the sounds on the MOS scale. The second listening test was a preference test where the listeners heard vowels corresponding to the same three processing types and they were asked to select which one they prefer to listen. A total of 10 listeners participated in the listening tests.

Figure 7 shows the results of the MOS test. The data indicates that the proposed system has a mean MOS higher than 2.5 for all the vowels, which can be considered a good quality score for ES samples. Figure 8 shows the data of the preference tests by combining all the vowels. Also these data indicate that the proposed method has succeeded in enhancing the quality of the ES vowels.

4. Conclusion

An enhancement system for ES vowels was proposed based on using a natural glottal pulse combined with second order polynomial Frequency Warping Function. A preliminary evaluation of the system was carried out on early stage Spanish ES vowels by comparing the system performance with a known reference method. Results obtained with a MOS evaluation show clear improvements in speech quality both in comparison to the original ES vowels and to sounds enhanced with the reference method. The good performance was corroborated with a preference test indicating that in the vast majority of the cases, listeners preferred to listen to the sounds enhanced by the proposed method. Future work is needed to study the system together with advanced stage ES speakers.

5. Acknowledgements

Special thanks to all my colleagues at Aalto University for their valuable support and time.

6. References

[1] G. Fant, "Acoustic theory of speech production." Mouton, The Hauge, 1960.

[2] Q. Yingyong, W. Bernd, and B. Ning, "Enhancement of female esophageal and tracheoesophageal speech," *Acoustical Society of America*, vol. 98(5, Pt1), pp. 2461–2465, 1995.

[3] Y. Qi, "Replacing tracheoesophageal voicing source using lpc synthesis," *Acoustical Society of America*, vol. 5, pp. 1228–1235, 1990.

[4] R. Sirichokswad, P. Boonpramuk, N. Kasemkosin, P. Chanyagorn, W. Charoensuk, and H. H. Szu, "Improvement of esophageal speech using lpc and lf model," *Internation Conf. on Biomedical and Pharamaceutical Engineering 2006*, pp. 405–408, 2006.

[5] M. Kenji, H. Noriyo, K. Noriko, and H. Hajime, "Enhancement of esophageal speech using formant synthesis," *Acoustic. Sci. and Tech.*, pp. 69–76, 2002.

[6] M. Kenji and H. Noriyo, "Enhancement of esophageal speech using formant synthesis," *Acoustics, Speech and Signal Processing, International conf.*, pp. 81–85, 1999.

[7] R. H. Ali and S. B. Jebara, "Esophageal speech enhancement using excitation source synthesis and formant structure modification," *SITIS*, pp. 615–624, 2006.

[8] K. Doi, H.and Nakamura, T. Toda, H. Saruwatari, and K. Shikano, "Statistical approach to enhancing esophageal speech based on gaussian mixture models," *Acoustics Speech and Signal Processing(ICASSP), 2010 IEEE International Conference*, pp. 4250–4253, 2010.

[9] O. Ibon, B. Garcia, and Z. M. Amaia, "New approach for oesophageal speech enhancement," *10th International conference, ISSPA*, vol. 5, pp. 225–228, 2010.

[10] B. Garcia and A. Mendez, "Oesophageal speech enhancement using poles stablization and kalman filtering," *ICASSP*, pp. 1597–1600, 2008.

[11] B. Garcia, I. Ruiz, A. Mendez, and M. Mendezona, "Oesophageal voice acoustic parameterization by means of optimum shimmer calculation," *WSEAS Trasactions on Systems*, pp. 489–499, 2008.

[12] R. Ishaq and B. G. Zapirain, "Optimal subband kalman filter for normal and oesophageal speech enhancement," *Bio-Medical Materials and Engineering*, vol. 24, pp. 3569–3578, 2014.

[13] R. Ishaq, B. G. Zapirain, M. Shahid, and B. Lovstrom, "Subband modulator kalman filtering for signla channel speech enhancement," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013.

[14] R. Ishaq and B. G. Zapirain, "Adaptive gain equalizer for improvement of esophageal speech," in *IEEE International Symposium on Signal Processing and Information Technology*, 2012.

[15] A. Suni, T. Raitio, , M. Vainio, and P. Alku, "The glottalHMM entery for blizzard challenge 2011: Utilizing source unit selection in hmm-based speech synthesis for improved excitation generation," in *in Blizzard Challenge 2011, Workshop, Florence, Italy*, 2011.

[16] T. Raitio, A. Suni, J. Yamagishi, H. Pulakka, J. Nurminen, M. Vainio, and P. Alku, "HMM-based speech synthesis utilizing glottal inverse filtering," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, pp. 153–165, 2011.

[17] T. Raitio, A. Suni, H. Pulakka, M. Vainio, and P. Alku, "Utilizing glottal source pulse library for generating improved excitation signal for hmm-based speech synthesis," in *IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP)*, 2011.

[18] P. Alku, "Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering," in *Speech communication*, vol. 11, no. 2, 1992, pp. 109–118.

[19] J. H. Chen and A. Gersho, "Adaptive postfiltering for quality enhancement of coded speech," *Speech and Audio Processing, IEEE Transactions on*, vol. 3, pp. 59–71, 1995.