Glottal Waveform Model for Oesophageal Speech

John M. O' Toole and Begoña García Zapirain
DeustoTech-LIFE, University of Deusto, Spain
email: j.otoole@deusto.es and mbgarciazapi@deusto.es

***Abstract*—Oesophageal speech, a mode of speech for laryngectomees, is of low quality and intelligibility comparative to normal (laryngeal) speech. Understanding the signal differences between oesophageal and normal speech will help future oesophageal speech enhancement methods. We aim to produce a method to synthesise oesophageal speech using a simple source–filter model. In this paper, we fit a parametric glottal waveform model to speech samples in our oesophageal database. (This glottal waveform represents the source component in the source–filter approach.) We added coloured noise to the glottal waveform model to produce realistic sounding oesophageal speech. Our fitting error measure, a spectral distance measure, reduces for all tested speech samples when adding the coloured noise. Yet missing from our synthesised signal is the rough-sounding quality often present in oesophageal speech. This work represents the first steps in developing a method to synthesise oesophageal speech.**

I. Introduction

Oesophageal speech is a mode of speaking without a larynx. Laryngectomees, people who have had a larynx removed in surgery, are without a glottis and no longer have an airway from the trachea to the vocal tract. Thus the normal, or laryngeal, mode of speaking is not possible for laryngectomees. Oesophageal speech involves expelling air up through the oesophagus while vibrating the upper part of the oesophagus; this part of the oesophagus therefore acts as a neoglottis. Oesophageal speech has been described as rough, course, harsh, awkward, with low intelligibility comparative to normal (laryngeal) speech [1]–[3].

Signal processing methods can play an important role in improving the quality and intelligibility of oesophageal speech and thus improve the quality of life for oesophageal speakers. Existing methods have shown this potential, for example the methods in references [4]–[7]. Yet further progress is warranted to produce a suitable system capable of transforming oesophageal speech into normal speech without losing speaker identity or speech emotion.

Our goal is to synthesise realistic-sounding oesophageal speech. We believe that in doing so we can better understand the specific signal characteristics of oesophageal speech. This information will, in turn, help produce new speech enhancement methods. Another benefit to synthesising oesophageal speech is that we will be able to generate unlimited samples of (synthesised) oesophageal speech. Our approach is to convert normal speech to oesophageal-sounding speech, and thus a third benefit is that we will be able to use objective measures [8] to quantify the improvement for oesophageal speech enhancement methods. Objective measures could be used because we have the reference signal, in this case the normal speech signal.

This paper reports our work-in-progress to producing a method to synthesise oesophageal speech. We use the standard linear predictive coding (LPC) approach which is a simple source–filter model of the physical speech production system. The vocal tract is modelled by an autoregressive (AR) model and the glottis, or neoglottis, is modelled by a parametric glottal flow model. The aim is to quantify the parameters of oesophageal and normal speech, compare the differences between two parameter sets, and then develop a method to map the distribution of normal speech parameters to a distribution for oesophageal speech.

We started with a glottal flow model for this paper and fitted a two-parameter glottal model to both oesophageal and normal speech samples. The database used consisted of sustained vowels and words spoken in Spanish. The two parameters were fitted using nonlinear programming and optimised to minimise the Itakura–Saito spectral distance [9] function. We also added coloured Gaussian noise to the parametric model to enable a better fit to the data. The coloured noise was added to flatten the resultant spectrum of the glottal flow model. We concluded from informal listening tests that the glottal model plus coloured noise produced similar sounding speech samples although in some samples the roughness commonly associated with oesophageal speech was not present.

II. Methods

Linear prediction coding (LPC) is a speech (voice) coder method. The method is a simple model of the physical speech production system, using an autoregressive (AR) model to represent the vocal tract and a simple parametric model to represent the glottal waveform. The speech signal is segmented into frames, short-time epochs of speech, and the parameters for the vocal tract (filter) and the glottal waveform (source) are estimated in this analysis stage. For the synthesis stage, the source and filter are constructed from the parameters and the source is filtered to produce the synthesised speech [10].

Even though this source–filter approach is a linear time-invariant system and therefore a simplification of the real speech production mechanism, it has been successfully used as a voice coding method. The speech signal for a particular frame is thus represented as

$$s(n) = \sum_{l=1}^{P} a_l s(n-l) + e(n) \tag{1}$$

where $s(n)$ is the discrete speech signal, a_l are the AR coefficients, P is the order of the model, and $e(n)$ is the residual signal [10]. The residual signal $e(n)$ is a crude approximation to the glottal waveform because the interaction between the glottal flow and the vocal tract is more complicated than a linear time-invariant system. Yet for many applications this approximation is sufficient [10].

Once we have estimated the AR parameters, we can use the inverse filtering method to estimate the glottal waveform; that is, we use $e(n)$ as an estimate of the glottal waveform as

$$e(n) = s(n) - \sum_{l=1}^{P} a_l s(n-l). \tag{2}$$

In our analysis we used a frame size of 40 ms with a Hamming window and an overlap between successive frames of 50%. We estimated the AR coefficients using an AR model set to $P = 12$. After inverse filtering we obtained $e(n)$ and then used a threshold-based method to estimate the fundamental frequency [5].

A. Parameter Selection

LPC methods encode the source signal $e(n)$ by replacing it with a parametric model of $e(n)$. These models range in complexity from the simple, zero-parameter, impulsive–train model to the 10-parameter glottal flow derivative models with added aspiration noise [11]. For our study, we used the following 2-parameter glottal flow derivative model [12], [13],

$$g(n) = \begin{cases} 2cn - 3dn^2, & 0 \leq ON_0 \\ 0 & ON_0 < n < N_0 - 1 \end{cases} \tag{3}$$

where

$$c = \frac{27A}{4(O^2 N_0)}, \quad d = \frac{27A}{4(O^3 N_0^2)}.$$

The parameter A represents the peak amplitude of the glottal flow, as a ratio to N_0; parameter O represents the open quotient of the glottal source, again as a ratio to N_0; and N_0 is the (discrete) period of the frame under analysis. Note we use the glottal flow derivative and not the glottal flow to account for the lips radiation effect [10].

To estimate the parameters A and O, we fitted the model in (3) to the residual signal $e(n)$ using a spectral distance measure. We used the Itakura–Saito spectral distance measure [9]

$$\epsilon = \frac{1}{N} \sum_{k=1}^{L-1} \left(\frac{E(k)}{G(k)} - \ln \frac{E(k)}{G(k)} - 1 \right) \tag{4}$$

where $E(k)$ is the magnitude spectrum of residual signal and $G(k)$ is the magnitude spectrum of glottal waveform model.

We used nonlinear programming to find the minimum value of the error function ϵ within constrained ranges for the parameters O and A. The open-quotient parameter O was initialised to 0.6 and was restricted to the range [0.1,0.9]; the amplitude parameter A was constrained to $[A_{\text{init}}/20, 10A_{\text{init}}]$, where A_{init} is the initial value for A.

This initial parameter A_{init} was set using the approach described by Fu and Murphy [13]. This is as follows:

- convolve the residual signal $e(n)$ with a 7-point Blackman window;
- numerically integrate this smoothed residual function to obtain an estimate of the glottal flow signal, $U(n)$;
- let $A_{\text{init}} = U_{\max}/N_0$, where $U_{\max}$ is the maximum value of $U(n)$.

B. Adding Noise

When we inverse filter, in (2), we unintentional extract the spectral trend of the glottal flow derivative [14]. The consequence of this poor estimation method means our residual signal $e(n)$ has a flat spectrum. A better estimate of the glottal flow derivative would, however, contain a spectral trend sloping downwards from the lower to higher frequencies [9], [13].

To force the Rosenberg model to have a spectral flat response, we add coloured Gaussian noise (CGN) to the synthesised glottal model in (3). This Gaussian noise also provides a model for aspiration noise and any other noise source in the oesophageal glottal speech signal.

To colour the Gaussian noise we filtered white Gaussian noise. This filter $H(z)$ is the inverse of the estimated spectral trend of the glottal waveform. We estimated this glottal spectral trend by fitting a second-order AR model $G(z)$ to the glottal model and then inverted this filter so that $H(z) = 1/G(z)$.

III. Results

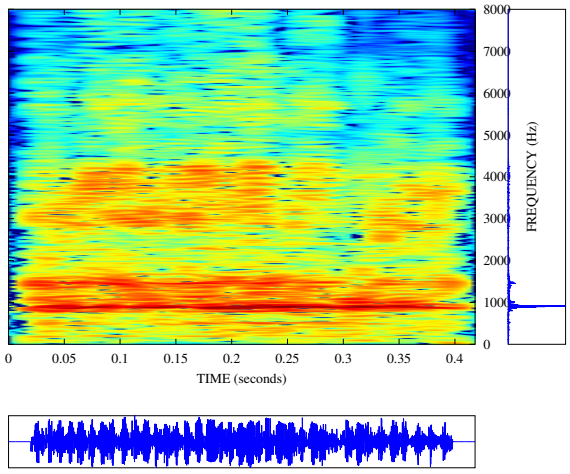
We fitted the glottal waveform model on a database of sustained vowels—*/a/, /e/, /i/, /o/*, and */u/* in Spanish—recorded from 4 oesophageal speakers and 4 normal (laryngeal) speakers. The histogram for the parameters A and O are in Fig. 1. Both parameter distributions appear similar—the open-quotient parameter O is concentrated around peaks in the region $[0.5N_0, 0.6N_0]$ and A is concentrated in the area < 0.005 for both normal and oesophageal speech samples.

To validate the usefulness of adding the CGN to the Rosenberg glottal model, we plot the error measure, the Itakura–Saito spectral distance measure, in Fig. 2. This figure shows that for all speech samples tested the error function decreased after the CGN was added, thus representing an improved fit for the model.

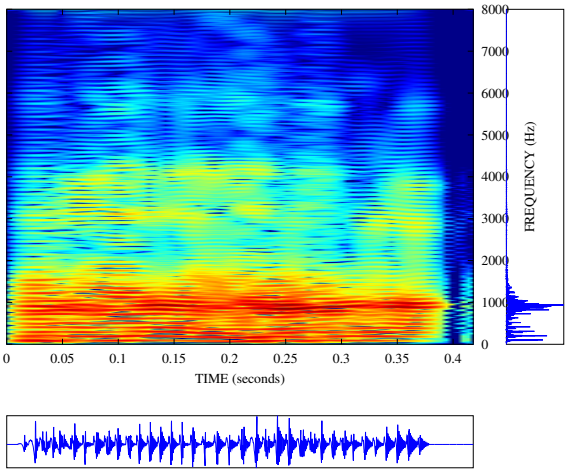
To show an example of the glottal waveform model we plot the time–frequency distributions (TFD) in Fig. 3. We use a separable-kernel TFD and not the spectrogram here to achieve finer time–frequency resolution [15]. We computed a decimated TFD using the algorithms in [16], [17].

IV. Discussion

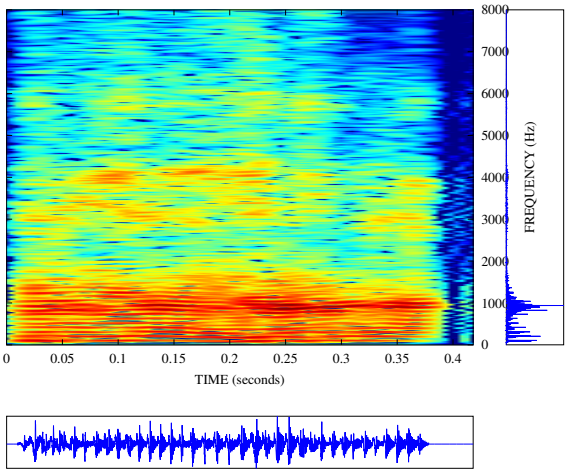
Informal listening tests indicate that the synthesised speech using the glottal flow model plus the CGN sounds closer to the original oesophageal speech sample comparative to the synthesised speech using the glottal flow model without noise. This observation is consistent with the error function measurements in Fig. 2. Although we note an improvement



(a) original sample



(b) glottal model without noise



(c) glottal model with noise

Fig. 3. Time–frequency distributions of oesophageal speech of vowel */a/*. The plots in (b) and (c) use the two-parameter Rosenberg model; the TFD in (c) uses Rosenberg model plus coloured Gaussian noise.

with the added CGN, our synthesised speech contains a larger degree of aspiration noise; more so than is apparent on the original (unprocessed) speech samples. Also, the synthesised speech is missing the rough quality common to oesophageal speech. We therefore need to investigate further to be able to reproduce this rough quality.

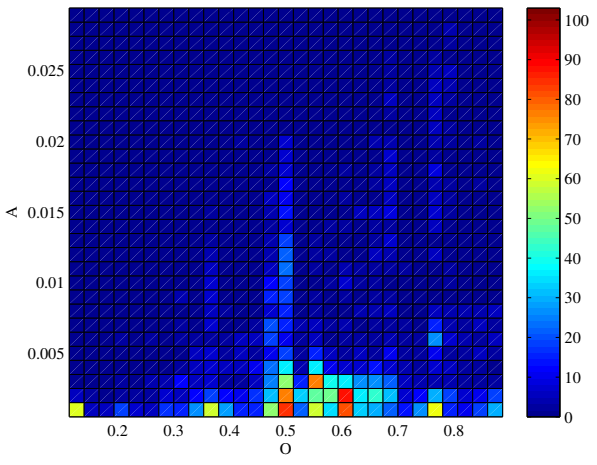
It is important to be aware of the limitations of our approach. The residual, obtained from inverse filtering, is a poor estimate of the glottal flow derivative signal. Yet there is probably little we can do to remedy this without an accurate measurement of glottal closure instants (which we did not have for our study). Of course, we need to be aware that the neoglottis of the oesophageal speaker is not the same as the glottis of normal (laryngeal) speakers. Thus even the concept of measuring glottal closure instants may not be suitable for oesophageal speakers. In fact, even the glottal waveform models may not be suitable for oesophageal speakers and better results may be obtained by deriving new models.

Another limitation of our approach is that our error function, the Itakura–Saito measure, ignores the phase of signal. This phase may be important for re-creating oesophageal sounding characteristics, as we know that phase contains significant information for speech signals [18], [19].

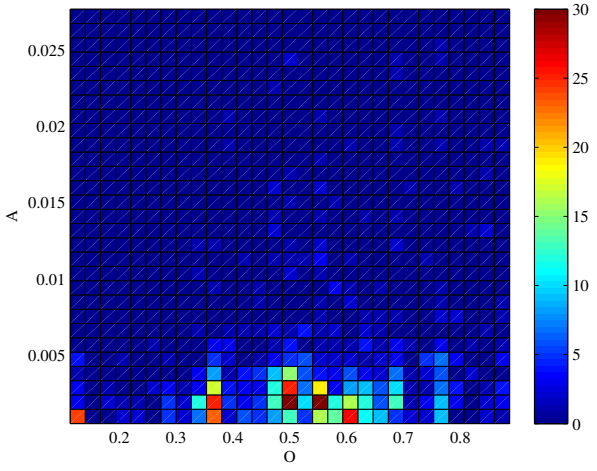
At this early stage we have little to conclude, but we now emphasise the importance of future work. First, we need to capture the rough-sounding characteristics of oesophageal speech into the glottal model. Next, we need to characterise the filter parameters for this source–filter approach. We will study the distribution and spectral characteristics of the noise of the line-spectral-pair coefficients which are used to model the vocal tract transfer function. Finally, we need to generate both the source and filter models, together, from normal (laryngeal) speech to produce the final synthesised speech.

REFERENCES

[1] J. K. MacCallum, L. Cai, L. Zhou, Y. Zhang, and J. J. Jiang, "Acoustic analysis of aperiodic voice: perturbation and nonlinear dynamic properties in esophageal phonation." *J. Voice*, vol. 23, no. 3, pp. 283–90, May 2009.



(a) normal speech



(b) oesophageal speech

Fig. 1. Histogram for the glottal parameters A and O using sustained vowels.

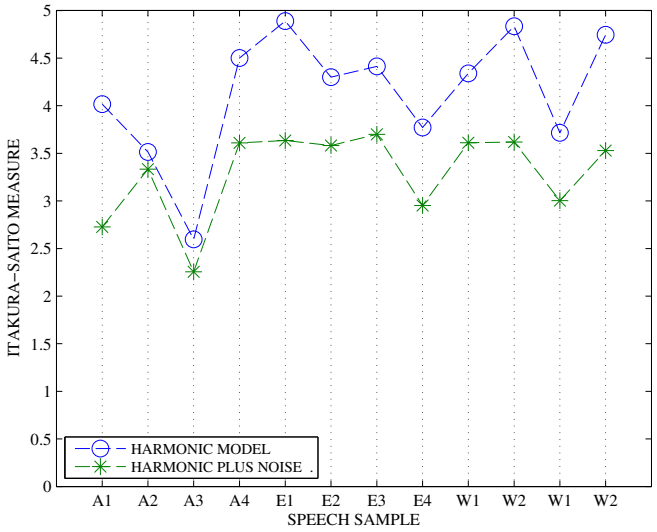


Fig. 2. Itakura–Saito spectral distance measure averaged over all frames for each speech sample in the database. The notation $A1 \ldots 4$ represents the */a/* vowel for four different oesophageal speakers; likewise $E1 \ldots 4$ represents the */e/* vowel for four different oesophageal speakers. The notation $W1$ and $W2$ represents the word *uno* spoken by two oesophageal speakers; $W3$ and $W4$ represents the word *cuatro* spoken by two oesophageal speakers.

[2] B. Garcia, I. Ruiz, and A. Mendez, "Oesophageal speech enhancement using poles stabilization and Kalman filtering," in *Proc. 2008 IEEE Int. Conf. Acoust., Speech, Signal Process.* Las Vegas, NV: IEEE, Mar. 2008, pp. 1597–1600.

[3] N. Yan, M. L. Ng, D. Wang, V. Chan, and L. Zhang, "Nonlinear Dynamics of Voices in Esophageal Phonation," in *33rd Annu. Int. Conf. IEEE-EMBS*, Boston, MA, 2011, pp. 2732–2735.

[4] Y. Qi, "Replacing tracheoesophageal voicing sources using LPC synthesis," *J. Acoust. Soc. Amer.*, vol. 88, no. 3, pp. 1228–1235, Sep. 1990.

[5] Y. Qi, B. Weinberg, and N. Bi, "Enhancement of female esophageal and tracheoesophageal speech," *J. Acoust. Soc. Amer.*, vol. 98, no. 5, pp. 2461–2465, Nov. 1995.

[6] K. Matsui and N. Hara, "Enhancement of esophageal speech using formant synthesis," in *Proc. 1999 IEEE Int. Conf. Acoust., Speech, and Signal Process.*, vol. 1, 1999, pp. 81–84.

[7] R. H. Ali and S. B. Jebara, "Esophageal speech enhancement using source synthesis and formant patterns modification," in *Signal Process. for Image Enhancement and Multimedia Process.*, ser. Multimedia Systems and Applications Series, E. Damiani, K. Yétongnon, P. Schelkens, A. Dipanda, L. Legrand, and R. Chbeir, Eds. Boston, MA: Springer US, 2008, vol. 31, pp. 279–288.

[8] P. Loizou, *Speech Enhancement: Theory and Practice (Signal Processing and Communications)*, 1st ed. CRC, 2007.

[9] M. Fröhlich, D. Michaelis, and H. W. Strube, "SIM—simultaneous inverse filtering and matching of a glottal flow model for acoustic speech signals," *J. Acoust. Soc. Amer.*, vol. 110, no. 1, p. 479, 2001.

[10] T. Quatieri, *Discrete-time speech signal processing: principles and practice*. Upper Saddle River, New Jersey: Prentice Hall, 2002.

[11] D. G. Childers, *Speech Processing and Synthesis Toolboxes*. Wiley, 1999.

[12] A. E. Rosenberg, "Effect of glottal pulse shape on the quality of natural vowels." *J. Acoust. Soc. Amer.*, vol. 49, no. 2, pp. 583–590, Feb. 1971.

[13] Q. Fu and P. Murphy, "Robust Glottal Source Estimation Based on Joint Source-Filter Model Optimization," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 2, pp. 492–501, Mar. 2006.

[14] A. Del Pozo, "Voice Source and Duration Modelling for Voice Conversion and Speech Repair," Ph.D. dissertation, University of Cambridge, 2008.

[15] B. Boashash, Ed., *Time–Frequency Signal Analysis and Processing: A Comprehensive Reference*. Oxford, UK: Elsevier, 2003.

[16] J. M. O' Toole, "Discrete quadratic time–frequency distributions: definition, computation, and a newborn electroencephalogram application," Ph.D. dissertation, School of Medicine, The University of Queensland, Nov. 2009. [Online]. Available: http://espace.library.uq.edu.au/view/UQ:185537

[17] J. M. O'Toole and B. Boashash, "Fast and memory-efficient algorithms for computing time–frequency distributions," *under review*, p. 13, 2011.

[18] K. Paliwal and L. Alsteris, "Usefulness of phase spectrum in human speech perception," in *Proc. Eurospeech*, 2003, pp. 2117–2120.

[19] A. Oppenheim, "Importance of Phase in Signals," *Proc. IEEE*, vol. 69, no. 5, pp. 529–541, 1981.