Quantifying Parameters of a Source–Filter 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*—Signal processing methods can improve the quality and intelligibility of oesophageal speech. Current methods show only moderate improvement leaving potential for better results. Quantifying parameters of oesophageal speech relative to laryngeal (normal) speech would help in the design of future enhancement methods for oesophageal speech. We quantified parameters of a source–filter model on a database of sustained vowels in Spanish from 4 oesophageal and 4 normal speakers. A ten-parameter glottal waveform model was used as the source and an autoregressive model was used as the filter. Classification, using a log-spectral distance measure, showed that all oesophageal speech samples were classified as whisper voice types; a voice type with a signal to noise ratio of -20 dB. Filter parameters representing spectral amplitudes and bandwidths had a large degree of variation for oesophageal speech comparative to the degree of variation for normal speech (Brown–Forsythe test, $F < 0.001$). Source metrics, noise to harmonic ratio (NHR) and variation in fundamental frequency, were also significantly greater for oesophageal speech (t-test, $P < 0.001$). These results show a greater degree of nonstationarity, and a noisier glottal waveform, for oesophageal speech comparative to normal speech.**

I. Introduction

Oesophageal speech is a method of speech used by those without a larynx. The airway from the trachea is surgically closed from the vocal tract during a laryngectomy and oesophageal speakers must expel air up through the oesophagus to the vocal tract. Without vocal folds the typical periodic glottal waveform that defines voiced speech is absent, and the oesophageal speaker must learn to produce a similar waveform by vibrating the upper part of the oesophagus to mimic the vibrating glottis. Oesophageal speech has been described as rough, harsh, course, awkward, with low pitch, volume, and intelligibility comparative to normal (laryngeal) speech [1]–[3].

There are some reports in the literature quantifying the differences between normal and oesophageal speech. Early speech enhancement methods by Qi *et al.* [4], [5] showed that, comparative to normal speech, different parameters were needed to fit a source–filter to model to oesophageal speech. Others later found similar results using the same source–filter model approach [6], [7]. Some studies have used common speech measures, such as fundamental frequency, jitter, and shimmer to quantify a difference between normal and oesophageal speech [8]–[10]; and more recently, nonlinear measures have been used as discriminating features between the two speech types [1], [3]. Our goal is to build on this body of knowledge by quantifying the parameters of a source–filter speech model. Specifically, we want to know how best to model the oesophageal glottal waveform and the degree of nonstationarity in the filter parameters.

For the analysis in this paper, we used a database of sustained vowels, */a/, /e/, /i/, /o/*, and */u/* in Spanish, recorded from 4 oesophageal speakers and 4 normal speakers. To quantify the parameters of the method, we used a LPC (linear predictive coding) approach as a source–filter model. Specifically, the method used an autoregressive (AR) model of the vocal tract (filter) and a synthetic waveform to model the glottal waveform (source). The glottal waveform model was predefined into 8 different classes of voice types: modal, vocal fry, breathy, whisper, falsetto, and harsh [11]. Each glottal waveform class was fitted to the source speech signal and a log-spectral distance measure was used to assess a classification error. Other quantification analysis included assessing the variability of the fundamental frequency and noise to harmonic ratio (NHR) of the source; and assessing the variability of the line spectral pairs (LSPs) to quantify the spectral variation of the filter model.

Results showed significant differences for both source and filter parameters between normal and oesophageal speech. The LSP coefficients had a greater degree of nonstationarity: these coefficients varied more from one frame to the next for oesophageal speech comparative to the variation of the coefficients for normal speech ($F < 0.001$). Source parameters were also significantly different between the two speech types: fundamental frequency varied more for oesophageal speech comparative to the variation for normal speech and NHR was significantly larger; for both tests, $P < 0.001$. Whisper voice type gave the best spectral fit for all of the (source) oesophageal speech samples in our database; normal speech, in comparison, was represented by modal (50%), whisper (24%), and breathy (19%) voice types. Whisper voice type had the lowest signal to noise ratio (-20 dB) of all the voice types. Thus our results show a higher degree of nonstationarity, from LSP coefficients and fundamental frequency tests, and a noisier source signal, from NHR and voice type classification, for oesophageal speech comparative to that for normal speech. This analysis gives a clear picture of the differences between normal and oesophageal speech using the source–filter model approach and can be used to develop new oesophageal speech enhancement methods.

978-1-4673-0753-6/11/$26.00 ©2011 IEEE

II. Methods

LPC (linear predictive coding) is a source–filter method to represent speech with a small number of parameters thus enabling voice coding. The speech signal is divided into short overlapping frames, approximately 30 ms, and a linear time-invariant filter is fitted to each frame [12]. The role of this filter is to estimate the transfer function of the vocal tract and therefore estimate the format frequencies.

LPC methods use an auto-regressive (AR) model as the filter; thus

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

where $s(n)$ is the discrete speech signal within an analysis frame, a_l are the AR coefficients, P is the order of the model, and $e(n)$ is the residual signal. Thus $e(n)$ approximates the glottal waveform but this approximation is not exact because the relation between the glottal waveform and the vocal tract is not accurately modelled by a simple linear time-invariant system [12]. For many applications however this approximation is sufficient [12].

A. Filter Parameters

For our analysis we used an AR model of order 16, that is $P = 16$, with a 30 ms frame. Each frame was windowed with a Hamming function and the frame overlap was 50%. All speech samples were down-sampled to 8kHz, after the appropriate anti-aliasing filtering.

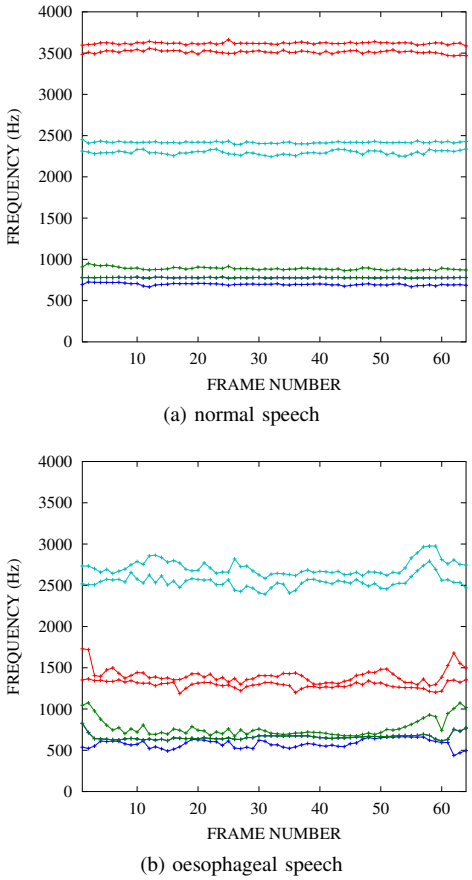
To quantify the differences in format frequencies we converted the AR coefficients to line spectral pairs (LSP) coefficients [13]. LSP coefficients are simply related to the frequency peaks of the signal [13].

We used only the 4 most significant LSPs, as we expect these to represent the most dominant spectral peaks. To determine the most significant LSPs, we first calculated the distance (difference) between the pairs and then selected the first 4 pairs with the smallest distance. We use this procedure because the distance between LSPs is inversely proportional to the amplitude of the spectral component [13]. An example of LSPs for oesophageal and normal speech are in Fig. 1.

Using the LSPs we did the following two tests:

1) For each frame, the AR model estimated 4 sets of LSPs of the current speech sample; we then subtracted the mean from each set, so we are left only with the distribution of the variance of the LSP coefficients. Next, we averaged these 8 coefficients (each pair contains two coefficients) into one set to obtain a distribution of the variance and then averaged these sets overs the four speakers. For each vowel, we computed the distribution for both the oesophageal and laryngeal voices and then statistical compared the variances of the two distributions.
2) We repeated this process but this time with the difference between the LSP coefficients.

The first test quantifies the variation, about a mean value, of the spectral components; the second test quantifies the variation, about a mean value, of the amplitude of these spectral components.



(a) normal speech

(b) oesophageal speech

Fig. 1. Example of LSPs (line spectral pairs) for normal and oesophageal speech of /a/ vowel. Only 4 LSPs with the smallest distance between the pair are shown here. Note the larger degree of variability for the oesophageal LSPs compared to the variability for the normal LSPs.

B. Source Parameters

An important parameter of the glottal waveform is the fundamental frequency of the signal. The term fundamental is used because the signal is a harmonic signal for voiced speech. We compared the variation of fundamental frequency, about a mean value, between the normal and oesophageal speech samples. The procedure was as follows:

1) estimate the pitch using a threshold-based correlation method [5] from the residual signal $e(n)$;
2) subtract the mean and combine over the 4 speakers for each voice type;
3) and then statistical compare the variance of each voice type.

Another measure we quantified is the NHR (noise to harmonic ratio) [14]. This measure was originally proposed to quantify the level of hoarseness in speech. For our study, we used this measure to quantify the amount of noise in the glottal waveform. Thus, we measured the NHR on the residual signal $e(n)$ using the NHR method in [14]. We then statistically compared the difference in the NHR distributions between

voice type	t_c (%)	t_p (%)	t_e (%)	t_a (%)	jitter (%)	SNR (dB)
modal	58	41	55	4	2	40
vocal fry	48	59	2.7	72	10	20
breathy	46	66	2.7	77	5	20
whisper	50	80	8	100	2	-20
falsetto	50	80	8	100	2	50
harsh	25	30	1	50	10	10

TABLE I
PARAMETERS FOR GLOTTAL WAVEFORM MODEL FOR EACH SPEECH TYPE. THE TABLES OMITS 4 PARAMETERS RELATING TO HOW THE ASPIRATION NOISE IS ADDED TO THE SIGNAL; THESE EXTRA PARAMETERS AND MORE DETAILS OF THE MODEL ARE IN REFERENCE [11].

normal and oesophageal speech.

To further characterise the difference between oesophageal and normal speech, we fitted the residual signal to 8 pre-defined glottal waveform models. These pre-defined models represent 8 different voice types: modal, vocal fry, breathy, whisper, falsetto, and harsh [11]. Each voice type model has ten parameters to define the synthetic glottal waveform. This model consists of the 4-parameter glottal waveform model from reference [15] combined with jitter and a 5-parameter aspiration noise model [11]. The values of the parameters are in Table I and more details are in reference [11]. This glottal waveform uses the fundamental frequency as an input parameter [11], [15].

For each speech sample, at each frame, we constructed all 8 voice types using the estimated fundamental frequency at each frame. We then calculated the magnitude spectrum of the residual signal and averaged this spectrum for each speech type over all frames. This allows an estimate of the power spectral density (PSD) as we know that the residual signal is a noisy signal for the oesophageal speech samples. This PSD estimate is written as

$$E(k) = \sum_{f=1}^{F} |E_f(k)|^2$$

where $E_f(k)$ is the discrete Fourier transform of $e(n)$ at frame f for F frames in the speech sample. Similarly, for each speech sample we estimate the PSD for the 8 voice types; that is,

$$V_l(k) = \sum_{f=1}^{F} \left| V_l^f(k) \right|^2$$

where V_l^f is the discrete Fourier transform of the glottal waveform model at frame f for each voice type $l = 1, 2, \ldots, 8$ and $V_l(k)$ is PSD estimate. Nonstationary information in the signal will be lost by this PSD averaging procedure; what we fit therefore is the time-averaged spectral content of the speech and synthetic glottal waveform signal.

To assess which voice-type is the best fit, we used a distance error measure: the log-spectral distance [16]. The distance measure, for voice type l, is defined as

$$d_l = \sum_{k=k_1}^{N/2} |\ln E(k) - \ln V_l(k)|^2 \qquad (2)$$

where N is the length of the $E(k)$ and $V_l(k)$. To ignore the influence of any low frequency artefacts that could be easily removed without perceived loss of quality to the speech signal, k_1 was to set $f_1 N/F_s$ where $f_1 = 25$ Hz and F_s is the sampling frequency. Before calculating the distance measure d_l, we removed the trend from the spectral magnitudes $E_f(k)$ and $V_l^f(k)$. In our source–filter model, the trend of the speech spectrum is completely described by the filter. Thus we removed this trend to isolate the source model. We estimated the spectral trend by fitting an AR model of order 2 to the signal and then subtracted this AR signal; this resulted in flat PSD estimates.

The final pre-processing step before calculating d_l was to normalised both PSD estimates by replacing $E(k)$ with $E(k)/(\sum_{k=0}^{N-1} |E(k)|^2)^{1/2}$, and likewise for $V_l(k)$, to ensure that both PSD estimates in (2) have the same energy value.

III. RESULTS

A. Line Spectral Pair Coefficients and Fundamental Frequency

To quantify the difference in variance between the zero-mean LSP coefficients for normal and oesophageal speech, we used a heteroscedastic test to determine the statistical difference of variances. A Kolmogrovo–Smirnoff test showed that neither distribution, for normal and oesophageal speech, were normally distributed. For each vowel, the Brown–Forsythe test computed an F-score of less than 1×10^{-10}, indicating that the two variances are heterogeneous.

Similarly, the distributions for the difference between the LSPs were not normally distributed and were heteroscedastic, as the Brown–Forsythe test computed an F-score of less than 1×10^{-10} for each vowel. For these two LSP tests, the variance calculated for the oesophageal speech samples were larger comparative to the normal speech.

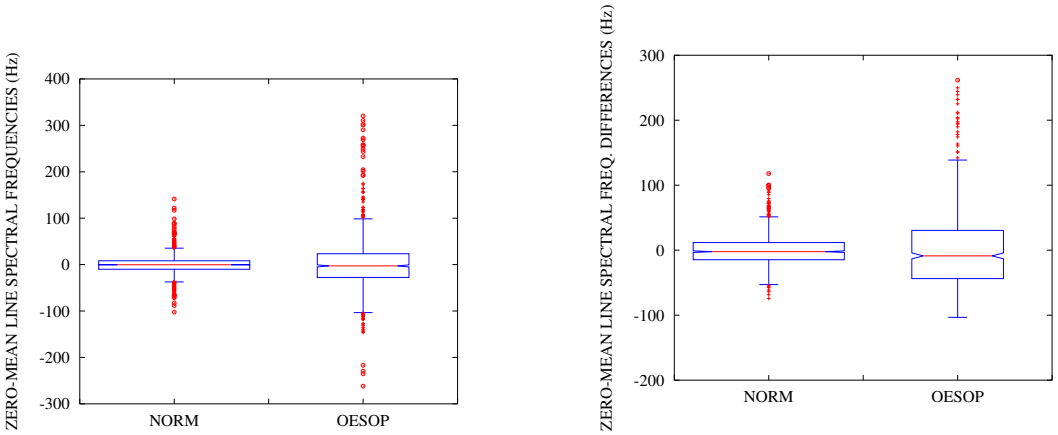
Fig. 2 shows the distributions of the LSPs and the difference between the LSPs for the */a/* vowel.

The next test showed that the variance of the fundamental frequency for oesophageal speech was (statistically) significantly larger compared to the variance of the fundamental frequency for normal speech. The F-score of the Brown–Forsythe test was less than 1×10^{-10} which indicates the variance between the normal and oesophageal speech distributions were statistically different. (We used the Brown–Forsythe test again as neither distributions were normally distributed.)

NHR for oesophageal speech was also significantly larger comparative to the NHR of the normal speech. Both distributions passed the Kolmogrovo–Smirnoff test and were therefore assumed to be normally distributed. Student t-test showed significant difference, with $P < 1 \times 10^{-16}$, between the two distributions. The distributions are shown in Fig. 3a

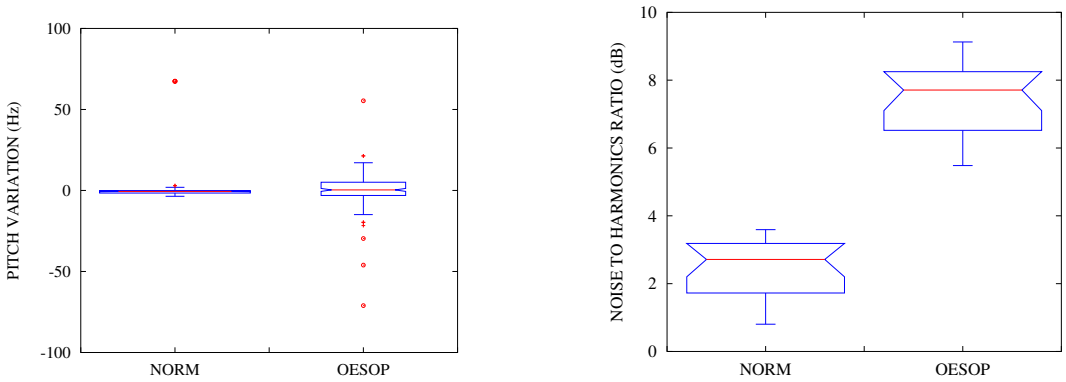
B. Classification of Glottal Waveform

The best class to represent oesophageal speech glottal waveform, out of the 8 classes in Table I, was the whisper voice type. For all the oesophageal speech samples, this whisper voice type had the lowest spectral distance measure. For



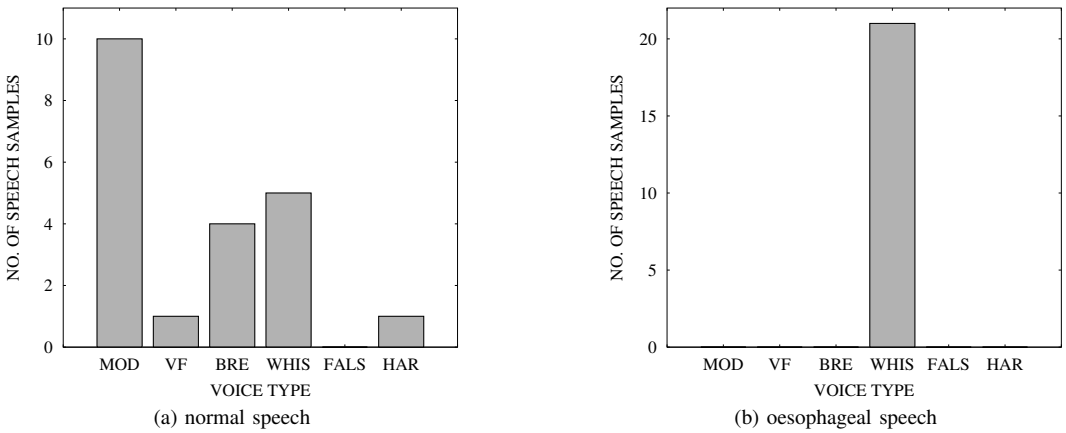
(a) zero-mean LSPs. (b) difference between zero-mean LSPs.

Fig. 2. Distribution of LSPs (line spectral pairs) for normal (NORM) and oesophageal (OESOP) speech for /a/ vowel.



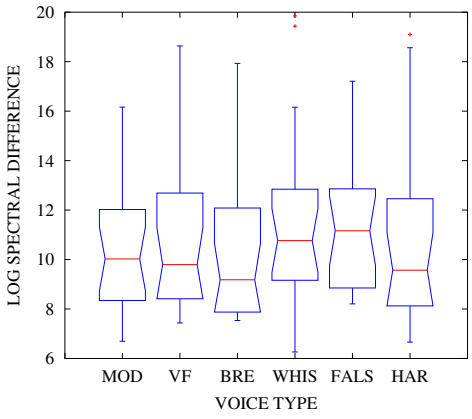
(a) variation of fundamental frequency (b) NHR (noise to harmonic ratio)

Fig. 3. Box-plot showing distributions for normal (NORM) and oesophageal (OESOP) speech.

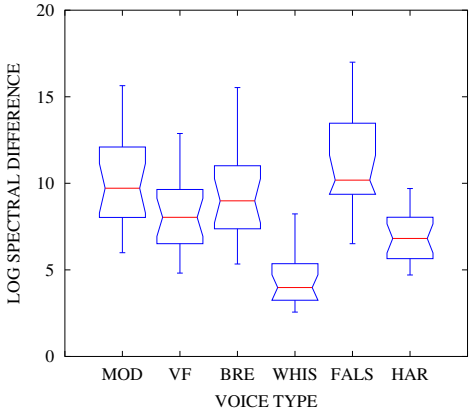


(a) normal speech (b) oesophageal speech

Fig. 4. Classification of voice types based on lowest spectral distance measure. Voice types are modal (MOD), vocal fry (VF), breathy (BRE), whisper (WHIS), falsetto (FALS), and harsh (HAR).



(a) normal speech



(b) oesophageal speech

Fig. 5. Box-plot of distribution of spectral distance measures for different voice types for normal speech samples. Voice types are modal (MOD), vocal fry (VF), breathy (BRE), whisper (WHIS), falsetto (FALS), and harsh (HAR).

normal speech, the classification results were more diverse. The most common voice type, with almost 50%, was modal; next, with 24%, was the whisper voice type; and third, with 19%, was the breathy voice type. These results are plotted in bar charts in Fig. 4.

The distributions of the distance measures for each class (voice type) are plotted in Fig. 5. The distributions in both plots, for normal and oesophageal speech, are not homoscedastic and therefore we can not apply an ANOVA test here. The picture, however, is consistent with classification results in Fig. 4: the whisper voice-type provides the best fit for oesophageal speech and the best fit for normal speech is distributed among the modal, breathy, and whisper voice-types.

IV. Discussion

For our database of sustained vowels the results showed that the parameters for modelling the vocal track were significantly different ($F < 0.001$) between the two speakers; both the change in LSPs coefficients and difference between LSP varied, over the analysis frames, more significantly for oesophageal speech compared with the variation over time for the normal (laryngeal) speech. We conclude that the filter component of a source–filter model for oesophageal speech has a larger degree of nonstationarity compared with the degree of nonstationarity for normal speech. A larger degree of variation implies a greater degree of time-varying change in frequency, or amplitude, or both frequency and amplitude, of the main formats. It remains to be seen, however, if this time-varying behaviour is caused by changes in the physical vocal tract after laryngectomy, or caused solely by the highly-variable glottal waveform input to this filter, or a combination of the both.

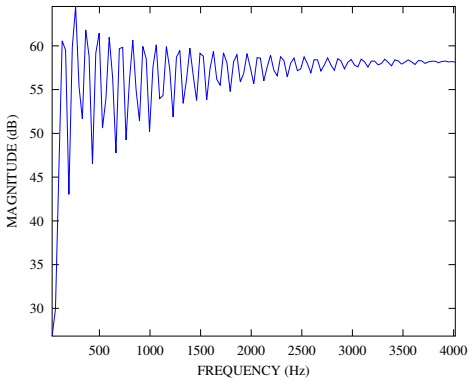
The results also showed that the variance of the fundamental frequency was larger for oesophageal speech compared to the variance of this frequency for normal speech. As the measure of variance of fundamental frequency is closely relate to jitter, these results are consistent with other studies comparing jitter in oesophageal and normal speech [1], [3], [10].

Similarly, the results showed larger values of NHR for oesophageal speech compared to values of NHR for laryngeal speech, where the NHR was calculated using the residual signal of the LPC approach. This finding is consistent with measures of NHR on the speech, not the residual, signal [1]. The implication of larger NHR for oesophageal speech is that either 1) the oesophageal voice is not capable of producing a glottal waveform with the same degree of harmonic structure as that for laryngeal speech or 2) that additional noise is added to the glottal waveform after a periodic signal is formed.

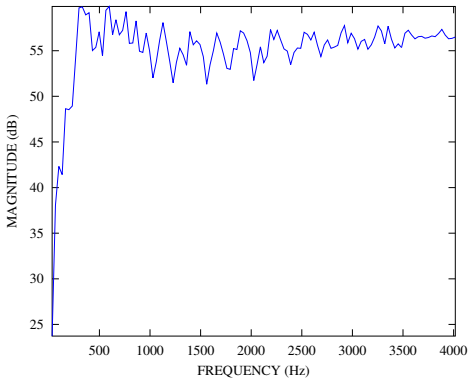
Supporting the proposition that the glottal waveform for oesophageal speech has little harmonic, periodic structure is the classification results: all oesophageal speech samples were classified as whisper voice types. The whisper voice type has the lowest SNR, at -20 dB, comparative to the other voice types, which are all above 20 dB.

Yet the variability in fundamental frequency could distort the classification results when averaged over many frames to produce the PSD estimate; thus, a similar result maybe be found for the residual signal with low NHR but with a large frame-to-frame fundamental frequency variation. Fig. 6 illustrates this effect of variable fundamental frequency on the PSD estimate. Because we found the oesophageal signal to have, comparative to normal speech, a large NHR measure we conclude that the residual signal for voiced oesophageal speech is a noisy signal with little periodic, and therefore harmonic, structure. This is consistent with our observations of oesophageal voiced signals in the time–frequency domain, as we found it hard to visually identify the usual harmonic patterns typically seen in normal (laryngeal) voiced speech.

A word of caution however: we need to careful in what we can conclude from fitting the residual LPC signal to voice types. First, the parameters of the model were fixed to give 8 classes of voice types; a better procedure would be to fit the parameters to each residual signal to produce a better fit. Second, we would expect modal speech to fit all normal speech samples but this was not a clear result and therefore we need to be careful in our interpretation of the voice types fitted to oesophageal speech. Third, the choice of distance measure can influence the classification results; a perceptual-based distance



(a) stable fundamental frequency



(b) variable fundamental frequency

Fig. 6. Example of PSD of synthetic glottal waveform, computed by averaging the magnitude spectrum over 30 frames. The PSD estimate is dependent on the fundamental frequency, which for normal speech is stable but for oesophageal speech is highly variable.

measure could give a measure closer to how we perceive the speech.

V. Conclusions

From our database of sustained Spanish vowels, we found statistical significance between all tested source and filter parameters for oesophageal speech comparative to the parameters for normal (laryngeal) speech. The glottal waveform for oesophageal speech is best described by a whisper voice type. These results should help inform future oesophageal speech enchantments methods that use the source–filter approach. Future work includes repeating the tests on a larger database with more speakers and to include unvoiced phonemes; assessing whether the nonstationarity of the filter component, a model of the vocal tract, is caused by, and if so to what degree, the nonstationary source; and developing an oesophageal model of speech to help construct and test speech enhancement methods.

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.

[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] T. Most, Y. Tobina, and R. Mimrana, "Acoustic and perceptual characteristics of esophageal and tracheoesophageal speech production," *J. Commun. Disorders*, vol. 33, no. 2, pp. 165–181, Apr. 2000.

[9] H. Liu, M. Wan, and S. Wang, "Acoustic characteristics of Mandarin esophageal speech," *J. Acoust. Soc. Amer.*, 2005.

[10] B. Garcia, I. Ruiz, A. Méndez, and M. Mendezona, "Oesophageal voice acoustic parameterization by means of optimum shimmer calculation," *WSEAS Trans. Syst.*, vol. 7, no. 5, pp. 489–499, 2008.

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

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

[13] I. V. McLoughlin, "Line spectral pairs," *Signal Process.*, vol. 88, no. 3, pp. 448–467, Mar. 2008.

[14] Y. Qi and R. E. Hillman, "Temporal and spectral estimations of harmonics-to-noise ratio in human voice signals," *J. Acoust. Soc. Amer.*, vol. 102, no. 1, pp. 537–543, Jul. 1997.

[15] G. Fant, J. Liljencrants, and Q. Lin, "A four-parameter model of glottal flow," *STL-QPSR, R. Inst. Technol. KTH*, vol. 26, no. 4, pp. 1–13, 1985.

[16] R. Gray, A. Buzo, A. Gray, and Y. Matsuyama, "Distortion measures for speech processing," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 28, no. 4, pp. 367–376, Aug. 1980.