DEVICE FOR PROSODIC SPEECH RESTORATION
A Multi-Resolution Approach for Glottal Excitation Restoration

O. Schleusing[1], R. Vetter[1], Ph. Renevey[1], J.-M. Vesin[2] and V. Schweizer[3]
[1]*Swiss Center for Electronics and Microtechnology, Neuchâtel, Switzerland*
[2]*Swiss Federal Institute of Technology, Lausanne, Switzerland*
[3]*University Hospital Lausanne, Lausanne, Switzerland*

Keywords: Speech processing, Speech restoration, Adaptive filters, Laryngectomy.

Abstract: This paper proposes a novel device for the restoration of authentic characteristics in pathological speech uttered by subjects with laryngeal disorders. The device acquires and analyzes the original speech signal and reconstructs a speech signal with improved, healthy-like features in real-time. The pathological excitation is replaced by concatenation of randomly chosen healthy reference patterns. To restore authentic features, intervals between subsequent reference patterns are obtained through a multi-resolution approach. Short-term pitch variability is reproduced through a statistical variation model. Middle-term pitch variability exploits the correlation at the middle-term time scale between pitch and signal envelope. Long-term variability is obtained through adaptive wavetable oscillators; a novel, reliable and computationally efficient method. Performance was assessed with respect to two authentic features, namely breathiness and prosody. Preliminary results have shown that breathiness of the restored signal is clearly reduced, while prosody related features are slightly improved.

1 INTRODUCTION

The degree of degradation found in pathological voices often engenders a decrease in a patient's speech intelligibility and thereby a severe limitation in its social oral interaction (Weinberg, 1986). For example, subjects who have undergone a laryngectomy suffer from degradation of their natural vocal excitation (Williams and Barber Watson, 1987; Most et al., 2000; Moerman et al., 2004). Laryngectomy is the common treatment after diagnosis of larynx cancer in an advanced stage and constitutes the partial or total removal of the larynx. This significantly reduces the patient's ability to produce voiced sounds due to the reduced or missing vocal cord functionality (van As, 2001; Pindzola and Cain, 1988). During speech rehabilitation the patient may learn to use an alternative voicing method, but the result usually is a noisy and intermittently obstructed voice that is not gender-discriminative due to its typically very low pitch. In accordance with the widely accepted source-filter-model in healthy speech processing (Fant, 1981), the vocal cords are essential since they provide an excitation signal with distinctive, periodical energy concentrations in the time domain. In contrast, the alaryngeal voice excitation consists of a flawed, distorted excitation signal where the glottal peaks are much less concentrated in the time domain. This results in an unpleasant and unnatural voice with an often intermitted and fluctuating periodicity. In addition, the speaker loses much of its control over pitch variability.

Several advanced voice restoration systems and methods have been presented in the past aiming at the improvement of the quality and intelligibility of the alaryngeal speech. In (Qi et al., 1995) methods based on linear prediction (LPC) for analysis and synthesis were used to enhance the perceived, subjective voice quality. In (Bi and Qi, 1997) modified voice conversion methods combined with formant enhancement were utilized to reduce the pathological speech signal's spectral distortions. In (del Pozo and Young, 2006) a voice restoration system is described that synthesizes speech from electroglottograph (EGG) pitch information and a jitter reduction model. In (Vetter et al., 2006) a system is presented that restores a pathological speech signal by replacing its pathological excitation with a concatenation of glottal reference patterns randomly chosen from a database extracted

Schleusing O., Vetter R., Renevey P., Vesin J. and Schweizer V. (2010).
DEVICE FOR PROSODIC SPEECH RESTORATION - A Multi-Resolution Approach for Glottal Excitation Restoration.
In *Proceedings of the Third International Conference on Biomedical Electronics and Devices*, pages 38-43
DOI: 10.5220/0002747500380043

from healthy speakers. The intervals between successive healthy glottal patterns are determined by the fundamental frequency (f_0) extracted from the original, pathological speech signal. Promising performances have been obtained in terms of reduction of breathiness and increase of the average pitch, but the resulting speech lacks authenticity due to the significantly reduced f_0-variability in pathological speech.

To improve these deficiencies, we propose a speech restoration device based on a multi-resolution pitch restoration method to increase the variability of the restored pitch in real-time. The long-term variability is deduced from the f_0-trend in the original speech signal. Middle-term variability is restored using the correlation between f_0 and the signal variance in natural speech. Short-term variability is restored using a statistical variation model and the signal envelope. The speech signal subsequently is reconstructed with the enhanced excitation and can be deployed in manifold applications such as voice enhancement systems or interactive support systems for voice rehabilitation and tutoring.

In the next Section we outline the pathological speech characteristics which lead to the development of our speech restoration system. In Section 3 we describe the multi-resolution pitch restoration method used by the proposed device. The results obtained by subjective listening tests are then presented and discussed in Section 4.

2 CHARACTERISTICS OF PATHOLOGICAL SPEECH

During laryngectomy, the larynx including the vocal cords and the laryngeal muscles are partially or totally removed (van As, 2001). Generally, postlaryngectomy patients may regain means of verbal communication in two ways. On one hand there exist electromechanical devices like the electrolarynx that inject vibrations into the trachea when held against the neck. The achievable voice quality of this approach is rather low since there is no intuitive control over voice quality or fundamental frequency. Yet, this method is attractive due to its simplicity and short learning phase. On the other hand, postlaryngectomy speakers may learn to use other tissues to substitute the functionality of the vocal cords. This way, the speaker retains, to some extent, intuitive control over aspects of the speech that allow the expression of prosody such as variations of the fundamental frequency. Unfortunately, the aptitude of the remaining tissue to produce a rich, harmonic sound is very limited. Its physical properties vary greatly among speakers and differ significantly from those of the vocal cords.

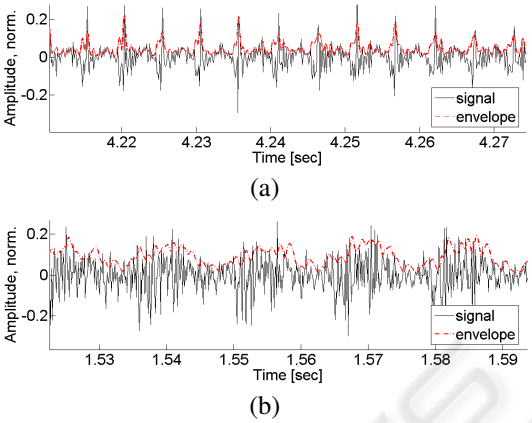

Figure 1: Segments of a laryngeal (a) and an alaryngeal (b) voice excitation signal and their respective envelopes.

In comparison to pathological voices, the aerodynamic-myoplastic processes in voiced, healthy speech production are relatively well studied (van As, 2001). Subglottal air pressure leads to a sudden, non-symmetric opening of the vocal cords and a release of this pressure. Subsequently, the musculature surrounding the glottis may force a closure of the vocal cords and the process starts all over. Varieties in glottis shape and size amongst humans lead to speaker-specific patterns for the opening and closing process as well as to the introduction of jitter in the period between subsequent glottal cycles. These effects amongst others lead to speaker specific voice characteristics.

Alaryngeal voice characteristics have been found to differ remarkably from that of healthy voices. Among subjects the position and shape of the neoglottis vary significantly (Qi et al., 1995). Often incomplete glottal closure can be observed. Furthermore, the flexibility and controllability of the neoglottis lacks greatly when compared with a healthy glottis, especially due to the absence of the laryngeal musculature. The mass of the neoglottis and resistance to mucus aggregation influence the absolute value and stability of the fundamental frequency in a disadvantageous manner. The alaryngeal oscillator tends to break down intermittently (Kasuya et al., 1986). Eventually, the resulting voice has an unnaturally low and instable pitch and often is found to have a hoarse, croaky and breathy character (Verma and Kumar, 2005). Figure 1 depicts fractions of LPC-estimates of a laryngeal and an alaryngeal voiced excitation signal. This figure highlights the alteration of the produced harmonic excitation due to the changed physiologic conditions. The glottal wave patterns in the excitation of the healthy speaker are highly con-

centrated in the time domain, whereas the excitation of the alaryngeal speaker appears merely as a modulated noise signal.

3 MULTI-RESOLUTION VOICE RESTORATION

3.1 Method

A general overview of the method implemented in the device is depicted in Figure 2. The articulation information and the voice excitation are separated in a primer LPC-based signal analysis. The obtained excitation signal is then divided into voiced and unvoiced segments, since we are only interested in restoring the voiced excitation signal. To restore the speech signal, the voiced excitation segments are replaced by a signal, which is formed by concatenating glottal reference patterns. These reference patterns were previously extracted from healthy subjects and are randomly chosen from a lookup-table. The intervals between successive reference patterns determine the fundamental frequency of the reconstructed speech signal. The fundamental frequency in pathological speech is degraded in terms of variability and stability and thus insufficient for a successful restoration of an authentic speech signal. To increase authenticity, the intervals between subsequent glottal waves are obtained through a multi-resolution approach on three different time scales:

- Long-term pitch variations are restored with instantaneous frequency estimations obtained from the alaryngeal voice excitation with the adaptive wavetable oscillator method.
- Middle-term pitch variations are strongly related to prosody and are restored using the correlation between pitch variations and instantaneous signal energy (Rosenberg and Hirschberg, 2006).
- Short-term pitch variability is introduced to the extracted pitch values to adhere the presence of jitter in pitch as it is found in healthy speech.

The improved excitation signal then is recombined with the unmodified unvoiced speech segments and the estimated articulation information.

In the following we describe the methods used to estimate and reconstruct the pitch information on the different time scales. Subsequently we describe the evaluation of the restored pitch in terms of the reduced breathiness and increased prosody.

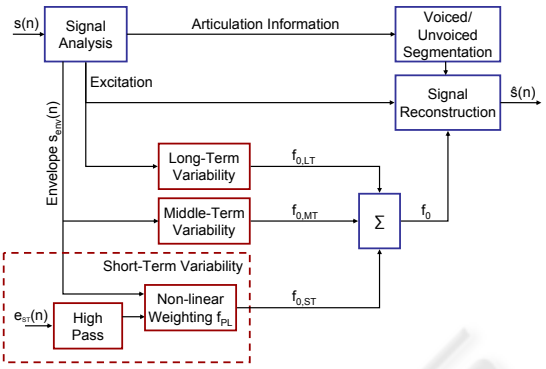

Figure 2: Block diagram of the multi-resolution pitch restoration method.

3.2 Long-Term Pitch Estimation

The selection of method for the extraction of the fundamental frequency of a specific signal depends on different characteristics of the signal itself:

- The signal's nature in terms of time-frequency distribution
- The amount and characteristics of additional harmful background noise
- The affordable computational complexity.

In general, pitch estimation methods can be classified into event detection methods and short-term averaging methods.Event detection methods such as for example zero-crossing (Gerhard, 2003) or threshold-guided maxima localization (Gerhard, 2003) are computationally inexpensive and yield high performance for well shaped signals in low-noise environments. Signals with higher harmonic complexity or increasing noise level require more advanced methods such as the matched filter method (Turin, 1960) or autocorrelation method (Un and Yang, 1977). They are based on short-term averaging and computationally more expensive. More advanced methods with yet increased computational complexity, decompose the signal into its Eigenspace components (Murakami and Ishida, 2001). Conjoint approaches (Mitev and Hadjitodorov, 2003) combine three different methods, namely in time, frequency and cepstrum domain.

For the device presented in this paper, the focus is clearly on the efficient utilization of the given computational resources. We propose to use a new pitch estimation method taking into account the demand for low computational load and for the pertinence and simplicity of fixed-point real-time implementation. The method is based on adaptive wavetable oscillators (AWO), a method recently published in (Arora and Sethares, 2007). An evaluation

of the method comparing it with other state-of-the-art methods for fundamental frequency estimation was presented in (Schleusing et al., 2009).

AWOs constitute a time-frequency method combining wavetables and adaptive oscillators. Wavetables generate periodic output signals by cyclic indexing of a lookup table that stores a single period of the waveform. Adaptive oscillators synchronize their output to both frequency and phase of the input signal. The indexing parameters of the AWO are determined by optimizing a well defined cost function such that the error between the wavetable output and an incoming, periodic signal is minimized.

The first step in the design of an AWO requires the selection of an appropriate pattern. This pattern should represent a high similarity with the signal pattern to be extracted and is stored in a wavetable as numerical, digital information. With respect to the above consideration, we use the energy distribution of the glottic excitation envelope as input (see Figure 1 (b)) and a Gaussian function as wavetable pattern. As one can observe, the envelope of energy during glottal patterns of the excitation signal has a high similarity with a Gaussian shape. A Gaussian function is easily controllable with only a few parameters such as a time index n, a phase offset in samples β and a temporal width σ:

$$w(n) = e^{-\frac{1}{2}\frac{(n-\beta)^2}{\sigma^2}} \tag{1}$$

Cyclic sampling is used to generate a periodical reference signal $v(n)$:

$$v(n) = w(k(n) \mod N) \tag{2}$$

where $k(n)$ is the cyclic sampling index

$$k(n) = (k(n-1)+\alpha) \mod N \tag{3}$$

$k(0)$ is initialized to 0, x mod N is the remainder operator, and α is the sampling step determining the sub-sampling rate of the wavetable pattern. The control parameters of the periodic output of Equation 1 are adaptively updated by using well understood gradient techniques (Haykin, 2001). The output of the wavetable oscillator thus is locked to the input signal and the parameter α is related to the fundamental frequency of the alaryngeal excitation signal by $\alpha/(NT_s)$, with T_s being the sampling period. The phase depends on the offset β of the sampling index. The adaption of the indexing parameters is achieved by minimizing a well defined cost function that gauges the error between the wavetable output and an incoming, periodic signal:

$$J(n) = s(n)v(n) \tag{4}$$

where $s(n)$ is the envelope of the extracted speech excitation signal.

Assuming that the phase and frequency of the input signal vary slowly over time one can follow these changes by moving the argument of the cost function slowly into the direction of the derivative:

$$\alpha(n+1) = \alpha(n) + \mu_\alpha \left.\frac{\partial J}{\partial \alpha}\right|_{\alpha=\alpha(n)} \tag{5}$$

and

$$\beta(n+1) = \beta(n) + \mu_\beta \left.\frac{\partial J}{\partial \beta}\right|_{\beta=\beta(n)} \tag{6}$$

It can be easily seen that the gradients $\frac{\partial J}{\partial \alpha}$ and $\frac{\partial J}{\partial \beta}$ are similar up to a constant.

$$\frac{\partial J}{\partial x} = \frac{\partial J}{\partial w}\frac{\partial w}{\partial x} \tag{7}$$

Indeed, they include the partial derivative of w, which is typically stored in a wavetable of N samples to minimize the computational load. The learning gains μ_α and μ_β should be chosen such that the oscillator can change rapidly enough to follow changes in the fundamental frequency and minimize noise influences. Since the adaptation of the frequency is much more sensible than that of the phase, μ_α should be much smaller than μ_β.

In (Schleusing et al., 2009) the AWO method was compared to several other methods such as the correlation method and the matched filter method. In an objective performance evaluation on synthetic and healthy speech signals with different amounts of additive white Gaussian noise (AWGN, ranging from 20dB to 0dB), the pitch estimation performance of the AWO method performed similar or better than the other methods. In a second, subjective evaluation, the methods were compared by reconstructing long-term trends of the fundamental frequency in an alaryngeal speech signal to a level, where the pitch quality is rated by listeners between fair and good. The performance of the AWO method was comparable to that of the other methods. Moreover, the AWO method has a very low computational complexity, which makes it a an ideal candidate for the proposed real-time speech restoration device.

3.3 Middle- and Short-Term Pitch Restoration

Middle-term pitch variability $f_{0,MT}$ is restored by exploiting the correlation between pitch and the signal envelope at this time scale. It was shown that prosody is strongly related not only to variations in pitch, but also to variations in the intensity of the speech signal (Rosenberg and Hirschberg, 2006). As can be observed in Figure 3, variations in the f_0 are directly

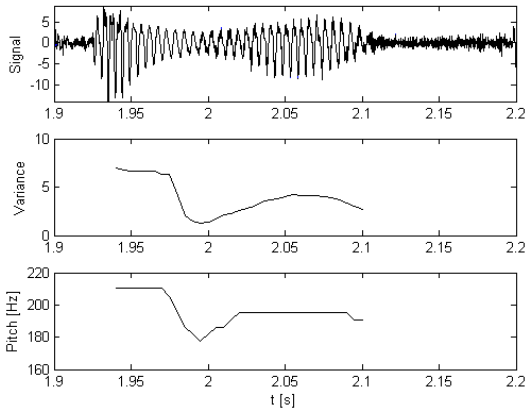

Figure 3: Correlation between speech signal envelope and pitch in a healthy speech signal.

related to variations of the signal envelope. In healthy speakers the f_0-variations may span several octaves, but pathological speakers have lost much of the ability to variate the fundamental frequency . Nevertheless, modulations of the intensity of the speech signal may allow a reconstruction of the f_0-variability due to their tight correlation. The key point of the presented method is to infer variations in pitch from variations in the signal envelope of the alaryngeal speech signal. Therefore, the segmental signal envelope is bandpass-filtered ($2-8\,Hz$) and then used to restore the middle-term pitch variability. This will give the pathological speaker a means to intuitively manipulate the pitch of the restored speech signal by manipulating the intensity of its speech.

Short-term pitch variability $f_{0,ST}$ is induced through high-pass-filtered ($f_c = 8\ Hz$) and weighted additive white Gaussian noise (AWGN). The weighting of this AWGN is determined through a signal-envelope-dependent nonlinear weighting inspired by recent findings in healthy subjects (Brockmann et al., 2008). In healthy voices, jitter in f_0 was found to be constantly low at sound pressure levels above $70-75\,dB$. Below this level though, jitter was found to steadily and sharply increase with falling sound pressure levels. We model this non-linear behavior with a piecewise linear function.

$$f_{PL}(s_{env}(n)) = \begin{cases} 0.1 & \text{if } s_{env}(n) \geq u \\ 0.1 + 0.9\frac{u - s_{env}(n)}{u-l} & \text{if } u > s_{env}(n) > l \\ 1 & \text{if } l \geq s_{env}(n) \end{cases} \quad (8)$$

where $s_{env}(n)$ is the logarithm of the averaged instantaneous signal envelope normalized with respect to the given acoustical configuration, u and l have been adjusted with respect to subjective listening tests.

4 RESULTS

An evaluation was performed to assess the successful restoration of authentic characteristics from pathological speech to a higher quality. A sustained sound of a vowel *a* of a pathological, male speaker with varying pitch was recorded at a sampling rate of $8kHz$ and 16 bits quantization. The speech signal was restored using the method as described in 3.1, implemented by the authors in the Matlab programming language (The Mathworks, 2006). Seven amateur listeners quantified the performance in terms of prosody and breathiness using a mean opinion score (MOS) by listening to the restored speech signals using consumer headphones. The relative contributions of short-term, middle-term and long-term pitch variabilities to the improved speech quality were assessed using three different system configurations. The pitch was restored from:

- LT: long-term pitch variability alone
- LT-MT: long-term pitch variability and middle-term variability of the signal envelope
- MR: multi-resolution approach, the long-term pitch variability, the middle-term variability of the signal envelope as well as the short-term variation of successive glottal waves.

The results displayed in Table 1 show that the proposed restoration approach improves performance with respect to both criteria. The contribution of the f_0-variability restored at the middle-term scale appears to be most significant (1.1 points compared to 0.1 points long-term variability alone). This seems to emphasize our assumption that signal energy variations at the middle-term scale can contribute to the restoration of prosody. The contribution of the MR approach yields no significant improvement to the perceived prosody. The high amount of standard deviation and the relatively small amount of listeners prohibits to draw general conclusions. Nevertheless, a positive trend can be recognized. Regarding the breathiness of the restored voice, a clear improvement (1.0 to 1.4 points) in all voices can be observed. For voices restored with the MR approach, the additional short-term variability seems to imply a degradation in term of increased breathiness. This could be due to the fact that short-term variability is related to the jitter of speech. Indeed, jitter may be perceived as a desired feature at a very low intensity level but becomes certainly harmful over a given threshold. This threshold depends on the subject's idiosynchrasies and may be adjusted to the laryngectomee's desire. We suggest that a more carefully designed non-linear model for the short-time variability contribution or a spectrally

Table 1: Mean Opinion Score (mean $\pm$ standard deviation) of 12 listeners assessing the quality of 3 different restored voices. Applied MOS-scale: highly improved-1, no improvement-3, highly degraded-5.

Method	Improved Feature	
	Prosody	Breathiness
LT	2.9 ± 0.7	1.7 ± 0.8
LT+MT	1.9 ± 0.7	1.6 ± 0.5
MR	2.0 ± 1.2	2.0 ± 1.2

shaped noise instead of the AWGN may reduce this undesired effect of the short-time pitch variability.

5 CONCLUSIONS

We presented a device for the restoration of authentic features in pathological voices. We have shown that the different methods utilized by the device can improve the prosody and breathiness of pathological voices to a different extent. Clearly, the study is limited by the small number of listeners and the small number of signals that the methods were applied to. Another limitation is the requirement of a relatively well developed pathological voice. In order to make the technology available for pathological speakers with less developed voices, additional signals have to be employed in future investigations. Nevertheless, the principal capability of the multi-resolution voice restoration device has been shown.

REFERENCES

Arora, R. and Sethares, W. A. (2007). Adaptive wavetable oscillators. *IEEE Trans. on Signal Processing*, 55 (9):4382–4392.

Bi, N. and Qi, Y. (1997). Application of speech conversion to alaryngeal speech enhancement. *IEEE Transactions on Speech and Audio Processing*, 5(2):97–105.

Brockmann, M., Storck, C., Carding, P., and Drinnan, M. (2008). Voice loudness and gender effects on jitter and shimmer in healthy adults. *Journal of Speech, Language and Hearing Research*, 51:1152–1160.

del Pozo, A. and Young, S. (2006). Continuous tracheoesophageal speech repair. *EUSIPCO*.

Fant, G. (1981). The source filter concept in voice production. *STL-QPSR*, 22:21–37.

Gerhard, D. (2003). Pitch extraction and fundamental frequency: History and current techniques. Technical report, University of Regina, CA.

Haykin, S. (2001). *Adaptive Filter Theory*. Prentice Hall.

Kasuya, H., Ogawa, S., Kikuchi, Y., and Ebihara, S. (1986). An acoustic analysis of pathological voice and its application to the evaluation of laryngeal pathology. *Speech Communication*, 5 (2):171–181.

Mitev, P. and Hadjitodorov, S. (2003). Fundamental frequency estimation of voice of patients with laryngeal disorders. *Information Sciences*, 156 (1-2):3–19.

Moerman, M., Pieters, G., Martens, J., van der Borgt, M., and Dejonckere, P. (2004). Objective evaluation of quality of substitution voices. *Eur Arch Otorhinolaryngol*, 261:541–547.

Most, T., Tobin, Y., and Mimran, R. (2000). Acoustic and perceptual characteristics of esophageal and tracheoesophageal speech production. *Journal of Communication Disorders*, 33(2):165–180.

Murakami, T. and Ishida, Y. (2001). Fundamental frequency estimation of speech signals using music algorithm. *Acoust. Sci. Technol.*, 22 (4):293–297.

Pindzola, R. and Cain, B. (1988). Acceptability ratings of tracheoesophageal speech. *Laryngoscope*, 98(4):394–397.

Qi, Y., Weinberg, B., and Bi, N. (1995). Enhancement of female esophageal and tracheoesophageal speech. *J. Acoust. Soc. of America*, 98(5 Pt 1):2461–2465.

Rosenberg, A. and Hirschberg, J. (2006). On the correlation between energy and pitch accent in read english speech. *Interspeech*, 1294-Mon2A3O.2.

Schleusing, O., Vetter, R., Renevey, P., Krauss, J., Reale, F., Schweizer, V., and Vesin, J.-M. (2009). Restoration of authentic features in tracheoesophageal speech by a multi-resolution approach. *Proc. of SPPRA 2009*, pages 643–042.

The Mathworks (2006). Matlab 2006b.

Turin, G. L. (1960). An introduction to matched filters. *IRE Transactions on Information Theory*, 6 (3):311–329.

Un, C. and Yang, S. (1977). A pitch extraction algorithm based on lpc inverse filtering. *IEEE Trans. ASSP*, 25:378–389.

van As, C. (2001). *Tracheoesophageal Speech: A multidimensional assessment of voice quality*. PhD thesis, University of Amsterdam.

Verma, A. and Kumar, A. (2005). Introducing roughness in individuality transformation through jitter modelling and modification. *ICASSP*, 1:5–8.

Vetter, R., Cornuz, J., Vuadens, P., Sola, I., and Renevey, P. (2006). Method and system for converting voice. European Patent. EP1710788.

Weinberg, B. (1986). *Laryngectomee Rehabilitation*, chapter Acoustical properties of esophageal and tracheoesophageal speech, pages 113–127. College-Hill Press, San Diego, CA.

Williams, S. and Barber Watson, J. (1987). Speaking proficiency variations according to method of alaryngeal voicing. *Laryngoscope*, 97(6):737–739.