Estimation of Vocal Fold Characteristics using a Parametric Source Model

Pavel Chytil[1,2] , Misha Pavel[1]

[1]Biomedical Engineering Department
Oregon Health & Science University, Portland, Oregon, USA
[2]Institute of Radioelectronics
Brno University of Technology, Brno, Czech Republic
{pchytil, pavel}@bme.ogi.edu

Abstract

We describe a new method to estimate vocal cord dynamics using a parametric model of glottis movements in order to assess the health of the vocal cords and detect pathological conditions of the larynx. The underlying model is based on that proposed by the Fujisaki-Ljungqvist, modified by reducing the number of parameters. The choice of a parametric model enabled us to assess the impact of the various model characteristics as well as the number of parameters. In the first step of the process we estimated the transfer function of the vocal tract and used it to approximate the model-generated speech. The general approach to the vocal tract estimation was based on the hypothesis that the vocal tract filtering action and the glottal forcing functions affect different, non-overlapping, frequency bands, and can therefore, be separated by homomorphic filtering. In the second step we used filtering of the acoustic waveform, using the estimated vocal tract transfer function $H(\omega)$ to estimate the shape of the glottal pulses. The actual estimates of $H(\omega)$ were obtained by applying an inverse filtering approach, using the cepstral method in conjunction with liftering (filtering) in the cepstral domain. The model parameters were estimated by maximizing the match between the model-generated and the observed speech signal. The general approach consists of finding the parameters of the FL model that would maximize the correspondence between the observed and synthetic utterances filtered by $H(\omega)$. The optimization was performed using the Nelder-Mead simplex search method because of the strong nonlinearities, discontinuities and the complex interactions among the model parameters. Prior to estimating the transfer function, the observed speech was filtered to remove the effects of lip radiation. In order to evaluate this approach we used the Kay Elemetrics Disordered Voice database, which comprises over 1,400 voice samples of approximately 700 subjects and includes sustained phonation and running speech samples from patients with a wide variety of organic, neurological, traumatic, and psychogenic voice disorders, as well as from 53 normal speakers. Estimated parameters of the FL model and its combinations were analyzed to determine the potential of this method to assess the health state of the speaker. We will illustrate the applicability of this technique to the problem of discriminating between healthy and pathological speech samples. The classification is based on the resulting estimates of the model parameters. The results suggest that the parameter estimates may provide a useful clinical tool for rapid unobtrusive, triage and diagnosis.

1. Introduction

The interaction of expiratory airflow with the vocal fold tissues is the primary source of human voiced sound production. The airflow through the larynx induces instability of the vocal folds. The resulting vocal fold vibrations modulate the airflow, giving rise to a periodic sequence of pressure pulses which propagates through the vocal tract and is radiated as voiced sound.

The characterization of vocal cords and their movements is an important factor in a number of application areas, including veridical speech synthesis (e.g., simulating a specific individual's speech) and for the assessment of vocal cords' health and the diagnosis of pathological conditions. In particular, there are a number of clinical conditions that affect directly or indirectly the physical properties of the vocal folds, and thereby, the pressure waveforms of the elicited sounds.

To date, such pathological conditions are usually detected by videostroboscopy using a laryngoscope specially designed to give detailed information regarding vocal fold anatomy and function. A long, thin scope is inserted through a patient's nose or throat and an image of the vocal tract is displayed on

Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. ISBN 0 9581946 2 9
University of Auckland, New Zealand. December 6-8, 2006. Copyright, Australian Speech Science & Technology Association Inc.

Accepted after abstract only review

a television screen. As the patient makes sound, the clinician can observe on the screen of a video monitor the relationship between the vocal positions and the resulting voice. This approach requires a well-equipped speech laboratory and a well-trained speech pathologist. Such laboratories are expensive to setup and operate, and therefore, available in only a limited number of clinics. An alternative examination based on the analysis of the acoustic waveform would enable earlier and more accurate detection in any clinical setting.

Current approaches to voice analysis, such as the estimation of jitter, shimmer and the harmonics-to-noise ratio are valuable for the characterization of regular phonation. However, in the case of a number of vocal cord dysfunctions and diseases in other regimes, the measurements are of limited relevance.

In this project we focused on the detection of diseases that affect the structure, and in particular those that alter the characteristics of the vocal folds. The ultimate objective of the computational approach described in this paper is to estimate the parameters of the glottal source model that would in turn enable the detection and classification of the glottis pathologies. Specifically, we describe an analysis of acoustic emissions that can be derived using a parametric model of the vocal cords.

Our starting point involves existing models of the vocal cords that resulted from extensive research in the field of speech synthesis, conducted for almost half a century. One of the first descriptions of the larynx, vocal tract and the pressure distribution appeared in late fifties (van der Berg, 1957), followed by a number of others (Flanagan & Landgraf, 1968; Ishizaka & Flanagan, 1972). The approach is based on a quasi-linear model of sound production in the vocal tract using the assumption that all voiced sound is generated at the vocal folds.

2. The glottal pulse model

There are several models that can be used to represent the glottal source waveform (Fujisaki and Ljungquist, 1986), but none of the methods of glottal source parameter extraction are restricted to any particular model. The Fujisaki-Ljungqvist model (FL) (Fujisaki & Ljungquist, 1986), which is a third-order polynomial approximation of the glottal flow derivative, has an advantage over the other models in that it is able to generate a slope discontinuity of variable height at glottal closure. We are using a modified model without slope discontinuity at the glottal opening. A stylized sample of the glottal pulse is shown in Figure 1.

The choice of a polynomial model provides a convenient way to vary the number of parameters, and thereby the level of detail in the modeling, which is convenient when evaluating the relative importance of the various parameters. In its most elaborate form, the model has three timing parameters controlling open phase duration, pulse skew and the time interval from glottal closure to maximum negative flow D, as well as three amplitude parameters controlling the slope at glottal opening A, the slope prior to closure B and the slope following closure C. Although the offset parameter A (see Figure 1) is not common in other models, we have included it since a secondary excitation is often noted at glottal opening. The rounded closure, which is often evident in the glottal flow waveforms, is sometimes attributed to a gradual glottal closure leaving a small residual flow after the main excitation stops. We consider that there is also a component attributable to a period of negative flow due to a lowering of the vocal cords following glottal closure.

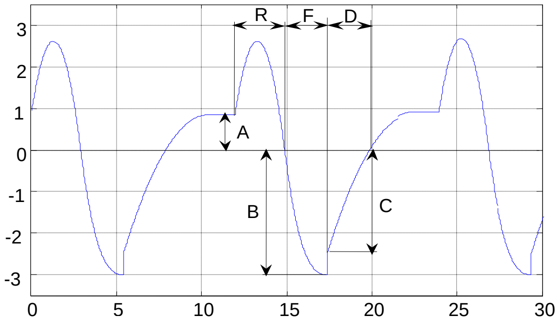
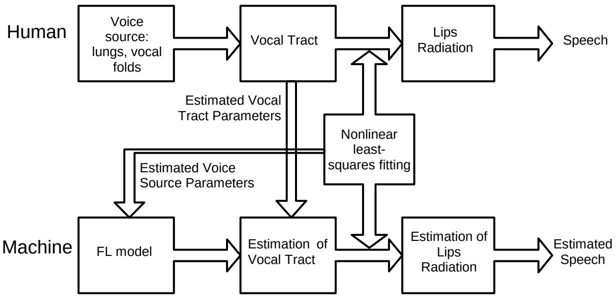


Figure 1: FL model showing parameters of the glottal pulse.

3. Parameter estimation

In order to characterize an individual's glottal pulses, we estimate the transfer function of the entire vocal tract and apply inverse filtering of the recorded wave form to obtain glottal pulses. Our parameter estimation procedure is based on inverse filtering using the cepstral method based on *liftering* (filtering) in the cepstral domain.



Human
Voice source: lungs, vocal folds
Vocal Tract
Lips Radiation
Speech
Estimated Vocal Tract Parameters
Nonlinear least-squares fitting
Estimated Voice Source Parameters
Machine
FL model
Estimation of Vocal Tract
Estimation of Lips Radiation
Estimated Speech

Figure 2: Block diagram of the parameters estimation process.

3.1. Algorithm description

A simple model of the speech generation and analysis processes is shown in Figure 2. The general approach consists of estimating the vocal tract transfer function *$H(\omega)$*, the pitch F_0, and then finding the parameters of the FL model that would maximize the correspondence between the observed and synthetic utterances filtered by *$H(\omega)$*. Prior to estimating the transfer function, the observed speech is filtered to remove the effects of lip radiation, i.e., it is convolved with

$$L(\omega) = 1 - \lambda e^{-j\omega}. \qquad (1)$$

In addition, the algorithm computed the pitch F_0 for all voiced speech segments using the cepstral method (Bostik &

Accepted after abstract only review

Sigmund, 2003). The knowledge of F_0 was used in conjunction with the FL model to generate the speech signals.

3.2. Vocal tract estimation using cepstral method

The general approach for vocal tract estimation is based on the notion that the frequency ranges of the vocal cord filtering action and the glottal forcing functions are not overlapping and uses homomorphic filtering whereby the multiplication of the transfer functions is transformed into an addition using logarithmic transformation. In particular, if we denote the Fourier transformation of the speech signal (after convolution with the inverse lip transfer function (1)) by X and the glottal pulse by G, then

$$X(\omega) = G(\omega)H(\omega), \tag{2}$$

and the power spectrum can be transformed to

$$C(\omega) = \log\|S(\omega)\|^2 = \log\|G(\omega)\|^2 + \log\|H(\omega)\|^2. \tag{3}$$

The logarithmic transformation underlying the cepstral analysis, in conjunction with a filtering in the cepstral domain, has been used to remove convolutional "noise" in a similar way that time domain filtering removes additive noise in the signal domain.

First, we apply an inverse filter for lip radiation as defined in (1)

$$X(\omega) = S(\omega)L(\omega)^{-1}. \tag{4}$$

The spectral representation of the air flow prior to passing through the lips can be described as a multiplication of the spectral representation of the glottal source component and the vocal tract transfer function

$$X(\omega) = DFT(w(t)x(t)) = H(\omega)G(\omega) = \|X(\omega)\|e^{j\Theta(\omega)}, \tag{5}$$

where $w(t)$ is a Hamming window. The resulting phase Θ, will be ignored during the cepstral analysis but will be used subsequently to reconstruct the generated speech signal. Cepstral coefficients $c(n)$ are obtained by taking the inverse DFT of $C(\omega)$ from (3).

$$\begin{aligned} c(n) &= IDFT(C(\omega)) = \\ &= IDFT(\log\|G(\omega)\|^2) + IDFT(\log\|H(\omega)\|^2). \end{aligned} \tag{6}$$

The additive representation together with the non-overlapping domains of the transfer functions permits us, in principle, to separate the vocal tract and voice source by *lifter*ing (filtering) of the cepstral coefficients in (6). This is due to the fact that cepstral coefficients $c(n)$, with lower index n characterizing the formant structure of speech (slow changes in the spectrum), the coefficients with higher index n characterizing glottal pulses (quick changes in the spectrum). For simplicity we used a rectangular window $l(n)$ with empirically selected length n_0 = 96 samples, corresponding to 2ms in the time domain using the sampling frequency of 48 kHz. The liftering process,

$$\tilde{c}(n) = l(n)c(n), \tag{7}$$

eliminates most of the spectral contribution of the vocal tract. Following the liftering process, the estimates of the vocal tract are obtained by reversing the process for computing cepstral coefficients. In particular, we estimated the vocal tract transfer function $\hat{H}(\omega)$ by

$$\hat{H}(\omega) = \left\|10^{DFT(\tilde{c}(n))}\right\|. \tag{8}$$

Although this estimate of the transfer function could potentially be used directly, in order to reduce the effects of noise, we regularized and smoothed the transfer function by representing it by a rational transfer function (ARMA). The coefficients of the polynomials A and B were estimated by minimizing the sum of the squared error between the empirical and the rational transfer functions, i.e.,

$$\min_{b,a}\left[\sum_{l=1}^{n}\left|\hat{H}(l) - \frac{B(\omega(l))}{A(\omega(l))}\right|^2\right], \tag{9}$$

where $A(\omega(l))$ and $B(\omega(l))$ are the Fourier transforms of the polynomials a and b, and n is length of the transfer function.

3.3. Empirical Evaluation: Dataset

We evaluated this approach using a database available from Kay Elemetrics (KayPentax, 1994). This database was originally assembled by the Massachusetts Eye and Ear Infirmary (MEEI) Voice and Speech Lab. The database comprises over 1,400 voice samples of approximately 700 subjects and includes sustained phonation and running speech samples from patients with a wide variety of organic, neurological, traumatic, and psychogenic voice disorders, as well as from 53 normal speakers. The speech samples were digitally recorded in a controlled environment using sampling frequency of 25 kHz or 50 kHz. For the purpose of this analysis, all utterances were resampled at rate 16 kHz to have all speakers at the same sampling frequency.

3.4. Fitting model to the data

The best-fitting set of six model parameters were estimated for each subject's data by maximizing the correlation at the point prior to the lip transformation. The optimization was performed using the Nelder-Mead simplex search method (Nelder & Mead, 1965), because of the complex error surface due to essential nonlinearities, discontinuities and the complex interactions among the model parameters.

For the purpose of this pilot evaluation we focused on three pathologies (A-P squeezing, gastric reflux and hyperfunction), which have the most occurrences in the database.

3.5. Linear Discriminant Analysis

We used linear discriminant analysis (LDA) as the initial approach to determine whether the estimated parameters of the model would enable us to distinguish between pathological and normal voice.

The LDA method is often used in pattern recognition application because of its simplicity and robustness. This method is based on finding the linear combination of features that best separates two (or more classes) of objects or events. Because of the “curse of dimensionality,” the optimal linear classifier is typically determined following a dimensionality reduction.

The starting point for the determination of the LDA is the selection of the feature vector $\vec{q}$ (also called observations, attributes, variables or measurements) for each sample from a known class *r*. This set of samples is called the training set. The classification problem is then to find a linear discriminant function

$$g_r(\vec{q}) = \vec{w}\vec{q} \tag{10}$$

that is maximum for the appropriate class r (Duda & Hart & Stork, 2000). LDA is based on the assumption that the probability density functions

$$p(\vec{q} \mid r = 1), \tag{11}$$

$$p(\vec{q} \mid r = 0), \tag{12}$$

are both normal, with identical full-rank covariances.

$$\sum\nolimits_{r=0} = \sum\nolimits_{r=1} = \sum \ . \tag{13}$$

In this case, the required probability $p(r \mid \vec{q})$ depends only on the dot product $\vec{w} \cdot \vec{q}$ where

$$\vec{w} = \sum\nolimits^{-1}\left(\vec{\mu}_1 - \vec{\mu}_0\right). \tag{14}$$

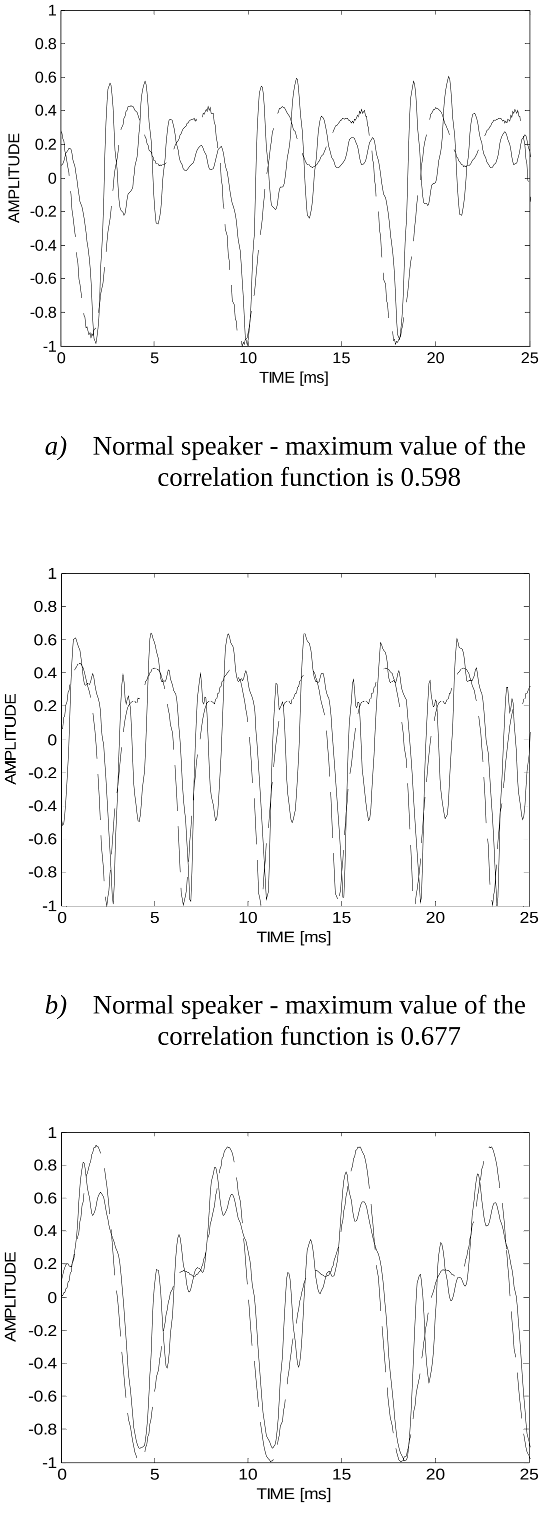
That is, the probability of an input $\vec{q}$ being in a class *r* is only a function of this linear combination of the known observations.

The vector of features used for the present classification problem included the maximum value of the correlation function, amplitude normalized parameters *A, B, C* of the FL model and parameters *D, F, R* of the FL model multiplied by F_0 of modeled subject.

4. Discussion and results

Sample of the results for three healthy and three pathological subjects are shown in Figure 3. Each plate in this figure represents the actual waveform for the subject measured at the lips and the best fitting waveform generated by the model. For each sample, we indicate the maximum value of the correlation function that represents the degree of correspondence between the model-generated and the observed speech sample; the value of unity would correspond to an ideal match.

A summary of the results of the fitting procedure is shown in Figure 4 using histograms of the resulting correlations for both groups of subjects. These results suggest that the model can account for a good proportion of the variance in both healthy and pathological subjects.

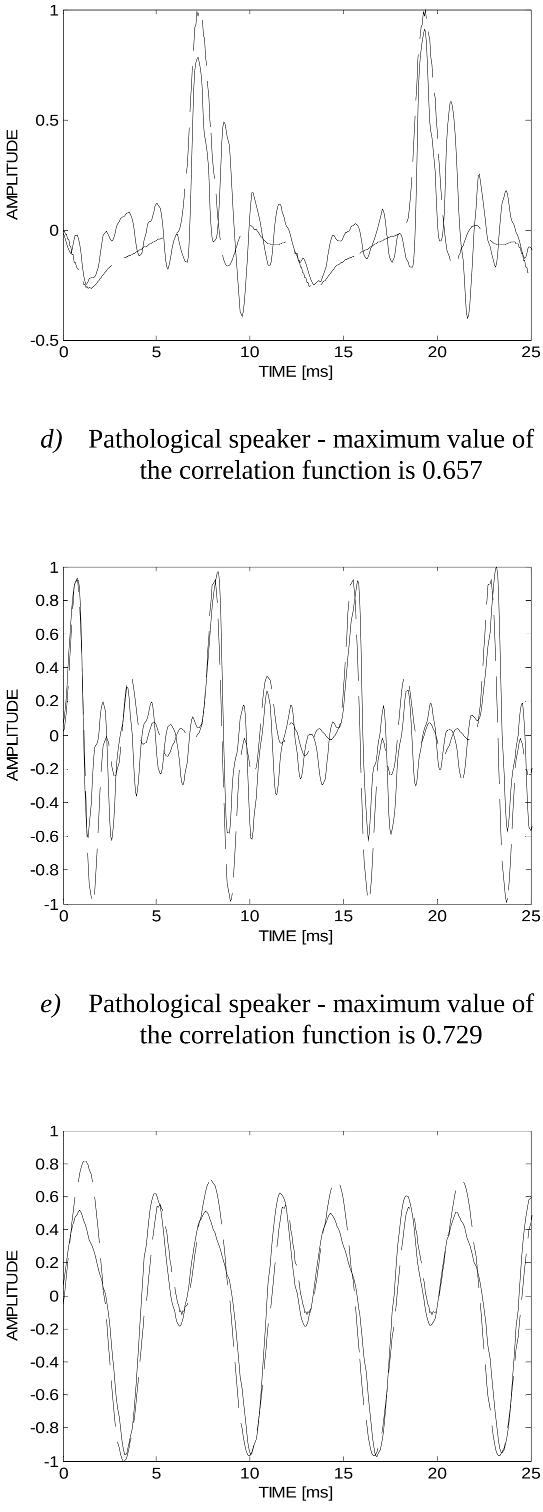


a) Normal speaker - maximum value of the correlation function is 0.598

b) Normal speaker - maximum value of the correlation function is 0.677

c) Normal speaker - maximum value of the correlation function is 0.896

Accepted after abstract only review



d) Pathological speaker - maximum value of the correlation function is 0.657

e) Pathological speaker - maximum value of the correlation function is 0.729

f) Pathological speaker - maximum value of the correlation function is 0.929

Figure 3: Signal fits for normal and pathological speakers, solid line is speaker and dashed line is synthesized modeled speech.

As expected there are a number of pathological subjects, for whom the model appears to be entirely inappropriate. There are several reasons for the poor match. One possible explanation is that the quasi-linear model of the sound production cannot account for the abnormal physical aspects of the dysfunction of the pathological speakers. However, another possibility is that the pathological speech is very unstable in both the fundamental frequency and in the shape of the glottal pulse. In those cases, the present approach based on the average pulse shape would not capture the speech generating process.

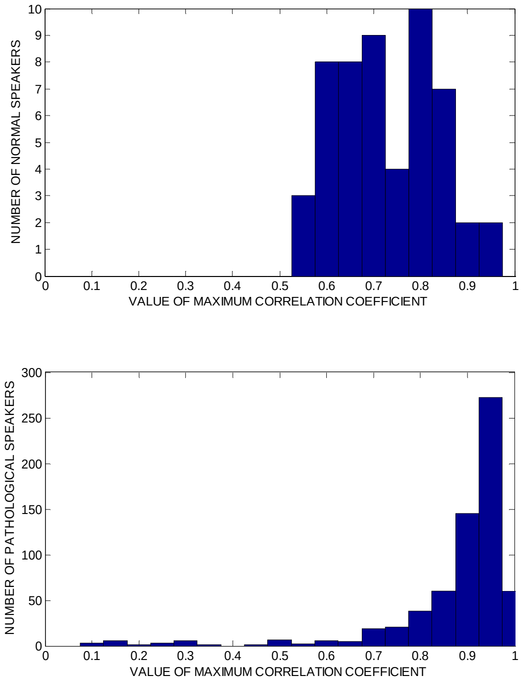


Figure 4: Histogram of the distribution of maximum correlation for normal and pathological speakers

The results of the binary classification process are shown in Table 1 in terms of proportion of the correct detection of pathological and normal cases (sensitivity and specificity). In each case the LDA was determined using binary classification of specific pathology vs. normal. The resulting performance of the glottal pulse model in conjunction with the simple LDA classification process is commensurate with many clinical tests.

Table 1: Results of positive detection of a health state

Health state	Proportion of Detection Pathological	Proportion of Detection Normal
A-P Squeezing	81%	80%
Gastric Reflux	80%	40%
Hyperfunction	100%	80%

Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. ISBN 0 9581946 2 9
University of Auckland, New Zealand. December 6-8, 2006. Copyright, Australian Speech Science & Technology Association Inc.

Accepted after abstract only review

5. Conclusion

This paper describes a novel approach to estimating the shape of the glottal pulses. The resulting estimates of the glottal pulses suggest that this may be a promising approach to characterizing the glottal sources for the purpose of speech generation. At the same time, these estimates may provide a powerful clinical tool for unobtrusive and rapid triage and diagnosis. Future work will be required to understand the relationship between the parameter values and the corresponding clinical pathologies.

6. Acknowledgement

This project was in part supported by a grant from the Intel Corporation. We thank Prof. Holly Jimison for her helpful suggestions.

7. References

Bostik, M. and Sigmund, M. (2003). *Methods for Estimation of Glottal Pulses Waveforms Exciting Voiced Speech*. In Proceedings of the 8th European Conference on Speech Communication and Technology, EUROSPEECH 2003, Geneva, 2389-2392.

Duda, R.O. and Hart, P.E. and Stork, D.H. (2000). *Pattern Classification* (2nd ed.). New York, Wiley-Interscience. ISBN 0-471-05669-3.

Flanagan, J.L. and Landgraf L. (1968). *Self-oscillating source for vocal-tract synthesizers*. IEEE Transactions on Audio and Electroacoustics, 16(1), 57–64.

Fujisaki, H. and Ljungqvist, M. (1986). *Proposal and Evaluation of Models for the Glottal Source Waveform.* In Proceedings of Acoustics, Speech, and Signal Processing, ICASSP’86, IEEE, Tokyo, Japan, pp. 1605-1608.

Ishizaka, K. and Flanagan, J. L. (1972). Synthesis of voiced sounds from a two-mass model of the vocal cords. *Bell System Technical Journal.*, 51, 1233–1268.

KayPentax (1994). *Disordered Voice Database.* ver 1.03. http://www.kayelemetrics.com/Product%20Info/CSL%20 Options/4337/4337.htm

Nelder, J.A. and Mead, R. (1965). A Simplex Method for Function Minimization*. Computer Journal*, 7, 308-313.

van den Berg, J.W. (1957). On the air response and the Bernoulli effect of the human larynx. *JASA*, 29, 626–631.