Estimating the voice source in noise

[1]*Gang Chen,* [2]*Yen-Liang Shue,* [3]*Jody Kreiman, and* [1]*Abeer Alwan*

[1]Department of Electrical Engineering, University of California, Los Angeles
[2]Dolby Australia
[3]Department of Head and Neck Surgery, School of Medicine, University of California, Los Angeles

gangchen@ee.ucla.edu,yshue@ee.ucla.edu, jkreiman@ucla.edu, alwan@ee.ucla.edu

Abstract

Estimation of the glottal source has applications in many areas of speech processing. Therefore, a noise-robust automatic source estimation algorithm is proposed in this paper. The source signal is estimated using a codebook search approach. The glottal area waveforms extracted from high-speed recordings of the glottis is converted to the glottal flow signals in order to evaluate the performance of the proposed source estimation algorithm. Results in clean and noisy conditions, on average, show that the proposed algorithm provides more accurate estimation than the software toolkit Aparat [1] as well as an earlier approach [2].

Index Terms: voice source, source estimation, speech analysis

1. Introduction

The voice source signal provides the excitation to the speech production system. The study of the voice source is important to many speech research disciplines, such as speech synthesis, voice quality analysis, and clinical assessment.

Many models have been proposed to represent the voice source, such as Rosenberg [3], Liljencrants-Fant (LF) [4], and Fujisaki-Ljungqvist [5] models. In our early study [6], LF model was used for source estimation. Some inconsistencies exist between the open quotient (OQ) estimated from the acoustic signal and OQ measured from high-speed imaging of the vocal folds, suggesting that a modification of the LF model may be necessary for accurately modeling the observed vibration of the vocal folds. A new source model was then proposed based on high-speed imaging of the larynx [2] in order to provide greater glottal pulse shape flexibility than the LF model. Results showed that the proposed model provided more accurate source estimation than LF model.

According to the linear speech production model [7], speech signals are generated by filtering the voice source by the vocal tract transfer function (VTTF). Generally, there are two types of approaches in source estimation. The first method relies on estimating the VTTF explicitly and then uses it to inverse-filter the speech signal. The residual signal obtained from inverse filtering is then fitted by a source model [8, 9, 10]. Inverse filtering typically requires estimating the formant frequencies explicitly. However, the widely used LPC-based formant trackers are known to be inaccurate for high-pitched phonations. Estimating the formant frequencies in noisy conditions remains far from robust. The inaccuracy in VTTF estimation would lead to inevitable inaccuracy in the source. In the second approach of source estimation, the voice source and the VTTF are estimated jointly and iteratively [11, 10], where the source estimation error due to the inaccurate VTTF estimation is compensated by searching a wide range of source-filter combinations. Synthesized speech and electroglottograph (EGG) signals recorded from natural speech have been used as references to evaluate the source estimation algorithms. However, the EGG signal is directly related to the contact area to the vocal folds, and thus does not provide an accurate shape of the glottal source signal.

In a recent study [2], glottal area waveforms obtained from high-speed recordings of the vocal folds were used as the reference to evaluate the source estimation algorithm. In that study, the glottal area was assumed to represent the glottal flow. However, since the production of glottal flow involves the interaction between the lung pressure and the glottal area function [12] as well as the interaction between the glottal area and the vocal tract system [13], the glottal area does not fully represent the glottal flow (e.g., the glottal flow pulse has a notable skewing rightward in time [14, 15]). The relationship between the glottal area and the glottal flow signal was quantitatively modeled using the three-mass vocal fold model in [16, 17]. In this paper, the glottal area obtained from high-speed imaging is converted to glottal flow using the three-mass model. The resultant glottal flow signal is used as the reference source signal to evaluate the accuracy of the proposed source estimation algorithm.

The source estimation method in [2] required estimating formant information. LPC-based formant estimators remain far from robust in noisy conditions, while manually-derived formants are impractical in applications. In this paper, a noise-robust automatic source estimation algorithm is proposed. This algorithm does not rely on explicitly estimating the formant frequencies to inverse-filter the speech signal. The source signal is estimated using a codebook search approach, and the method is a modified version of [2].

2. Data

The data used in this study are the same as those used in [2]. A brief summary of the data is as follows. Synchronous audio and high-speed video recordings of the vocal folds were collected from six subjects, three females and three males. None of the subjects had a history of voice disorders. Speakers were asked to sustain the vowel /i/ for approximately 10 seconds while holding voice quality, fundamental frequency (F_0), and loudness as steady as possible. Although the subjects were asked to produce the vowel /i/ for each recording, the vowel qualities were somewhat close to /æ/ or /ɛ/ due to the positioning of the laryngoscope. Manually measured formant frequencies were documented in [18]. Across tokens, speakers were asked to vary their F_0 (low, normal, and high) and voice quality (pressed, nor-

This work was supported in part by NSF grant IIS-1018863 and NIH/NIDCD grant DC01797.

mal, and breathy) quasi-orthogonally, resulting in nine recordings from each speakers. Details of F_0 values for each speaker can be found in [18]. For each recording, one second samples of audio and video were retained from the most stable sections for analysis. Gaussian white noise was also added to the audio signal to test the robustness of the source estimation algorithm. Three SNR levels were used: 20dB, 10dB, and 5dB.

High-speed imaging of the vocal folds were recorded at 3000 frames/second at a resolution of 512×512 pixels using a FASTCAM-ultima APX camera (Photron Ltd., San Diego). The glottal area was extracted from the first 150 image frames of each high-speed recording using a series of edge-detection and region-growing algorithms. The algorithm parameters were manually adjusted and glottal area segmentation was visually examined for each image to ensure accuracy. A detailed description of the algorithm can be found in [18]. Each cycle of the glottal area waveform was marked by locating the first instant of glottal opening, when glottal closure was complete. When the glottis did not close completely, the minimum glottal area points were recorded. The glottal area waveform was averaged across the glottal cycles to produce a single-cycle waveform which was representative of the 150 frames (50 ms) analyzed for that utterance. In order to evaluate the proposed source estimation method, the OQ was calculated from the averaged glottal area waveform as the time from the first opening instant to the onset of maximum closure (or minimum area), divided by cycle duration.

3. Method

3.1. Area to flow conversion

The glottal area extracted from high-speed images was converted to glottal flow by using the Matlab toolkit LeTalker [19]. LeTalker is a Matlab GUI version of the three-mass vocal fold model originally published in [16] and updated in [17]. Parameters such as muscle activation level and respiratory pressure can be specified as inputs to calculate the glottal area, the glottal flow, and the resultant speech signal. Both subglottal and supraglottal (vocal tract) systems are included to simulate their interactions with the vocal folds. In this work, the vocal tract shape was set to that of the vowel /i/ according to vocal tract area functions reported in [20] and all the other parameters were set to the default values in LeTalker when converting the glottal area to glottal flow.

Figure 1 shows an example of the glottal area extracted from high-speed recording and the resultant glottal flow calculated from LeTalker. As expected, due to the inertia of the air column [15], the glottal flow pulse is notably skewed rightward in time, as noted by previous researchers [14, 15].

3.2. The modified source model

The proposed model is a modified version of that proposed in [2]. The model has five parameters: the fundamental period (T_0), open quotient (OQ), asymmetry coefficient (α), speed of the opening phase (S_{op}) and speed of the closing phase (S_{cp}). An example of a model waveform is shown in Figure 2. t_o and t_c are the durations of the opening and the closing phase, respectively. Details about the derivation of the model and the parameters can be found in [18].

Using the notation from this figure, $OQ = \frac{t_0+t_c}{T_0}$, $\alpha = \frac{to}{to+tc}$, S_{op} is the waveform amplitude at the bisect instant of the opening phase, and S_{cp} is the waveform amplitude at the bisect instant of the closing phase. With the exception of T_0, the four other parameters all range from 0 to 1.

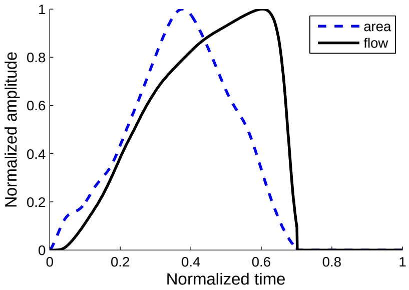

Figure 1: *An example of the glottal area extracted from high-speed images and the resultant glottal flow calculated using LeTalker*

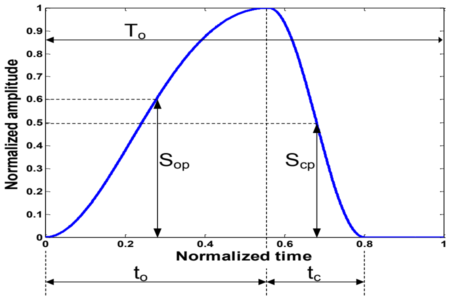

Figure 2: *Example of the proposed model with OQ=0.8,* $\alpha = 0.7$, $S_{op} = 0.6$, *and* $S_{cp} = 0.5$.

Mathematically the proposed model $u(t)$ is defined as:

$$u(t) = \begin{cases} f(\frac{t}{t_0}, \lambda_{Sop}) & 0 \leq t \leq t_0 \\ f(\frac{(t_o+t_c-t)}{t_c}, \lambda_{Scp}) & t_o < t \leq t_o + t_c \\ 0, & t_o + t_c < t \leq T_0 \end{cases} \tag{1}$$

where

$$\lambda = 12 \cdot (0.5 - S) \tag{2}$$

$$f(x, \lambda) = \frac{1}{\pi(e^{\lambda}+1)}\{e^{\lambda x}[\lambda sin(\pi x) - \pi cos(\pi x)] + \pi\} \tag{3}$$

λ is an intermediate slope parameter which controls the slopes of the waveform in the opening and the closing phase. λ_{Sop} and λ_{Scp} are the λ values when $S = S_{op}$ and $S = S_{cp}$ respectively. As shown in the equations above, given the five input model parameters (T_0, OQ, α, S_{op}, and S_{cp}), the intermediate slope parameter λ needs to be calculated in order to generate the output source waveform.

This modified model simplified the computational complexity of the model in [2] by redefining S_{op} and S_{cp} as amplitude domain measures. S_{op} and S_{cp} were originally defined as time domain measures in [2], where a time-consuming intermediate optimization step was required to calculate the slope parameter λ given $S = S_{op}$ or $S = S_{cp}$. By redefining the S_{op} and S_{cp} as amplitude domain measures in the modified model, an approximate trivial closed form solution of λ exists, as shown in Equation 2. The output source waveform can be calculated directly without the intermediate optimization step. The computation time of calculating the model waveform given the model parameters has been reduced by 90%, on average.

The model in [2] was derived from the glottal area data and was used for source estimation. In that study, the glottal area was assumed to represent the glottal flow, and the same glottal area data where the model was derived from was used to evaluated the source estimation algorithm. The modified model in

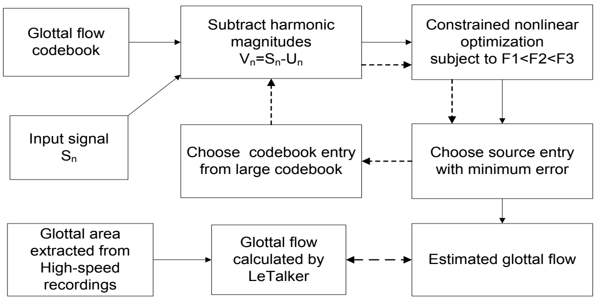
Glottal flow codebook
Subtract harmonic magnitudes $V_n = S_n - U_n$
Constrained nonlinear optimization subject to F1<F2<F3
Input signal S_n
Choose codebook entry from large codebook
Choose source entry with minimum error
Glottal area extracted from High-speed recordings
Glottal flow calculated by LeTalker
Estimated glottal flow

Figure 3: *Flowchart of the proposed source estimation algorithm*

this paper is also derived from the glottal area data and is used for source estimation. However, the glottal flow signals are used to test the source estimation method. The separation of the developing data and the testing data provides a basis where generalization of the source estimation method can be evaluated.

3.3. Source estimation procedures

The source estimation process is illustrated in Figure 3. In this method, a codebook is generated by the proposed source model. The harmonic magnitudes of the input acoustic signal are calculated and normalized to the first harmonic magnitude (the n-th normalized harmonic magnitude is denoted as S_n). The derivative of the source codebook entries are calculated to account for the radiation effect of the lips. The magnitudes of each codebook entry derivative are calculated in the same way (the n-th normalized harmonic magnitude is denoted as U_n). The vocal tract shape is obtained by subtracting the source harmonics from the acoustic signal harmonics ($S_n - U_n$). The residual signal is used for a constrained nonlinear optimization. A 3-formant VTTF is used here. In summary, for each of the entry in the source codebook, the following is performed:

$$\text{minimize} \quad E = \sum_{n=2}^{N} (S_n - U_n - V_n)^2 \cdot W_n \tag{4}$$

$$\text{subject to} \quad F_1 < F_2 < F_3 \tag{5}$$

where V_n is the n-th harmonic magnitude of the VTTF represented by three formants F_1, F_2, and F_3. Bandwidth values are based on the formant-bandwidth mapping formula in [21]. W_n is the weighting function and is empirically chosen as

$$W_n = \begin{cases} 2^{12-n} & 2 \leq n \leq 12 \\ 1, & n > 12 \end{cases} \tag{6}$$

The value of the error term E is recorded with the source entry. After searching the entire codebook, the source entry with the minimum error E is selected.

Note that in [2], formant information is required for source estimation as an input, while no explicit formant information is needed in the proposed approach. It is well known that formant estimation in noisy conditions remains far from robust and LPC-based formant trackers have deficiencies for high F_0 phonations. Thus, it is desirable to develop source estimation algorithms without relying heavily on the accuracy of formant estimation. The formant frequencies and the source signal are searched and evaluated jointly, rather than determining the formant frequencies to inverse-filter the speech signal. The optimal combination of the formants and the source signal is the final output. Although the recorded data only contain vowel /i/, the proposed algorithm is also suitable for source estimation for other vowels.

As in [2], two iterations of the search algorithm were used to reduce the computational complexity. The first iteration used a small codebook to search for the source parameters. The small codebook was generated by varying the OQ and α in the following way: OQ from 0.4 to 1.0 with an increment of 0.1; α from 0.5 to 0.9 with an increment of 0.1; S_{op} and S_{cp} were set to a constant value of 0.5. This generated a source codebook with 35 entries. The source entry selected from the first iteration was used for finding the final source entry in the second iteration from a larger codebook. The larger codebook was generated by the following setting: OQ from 0.35 to 1.00 at an increment of 0.01, α from 0.5 to 0.9 at an increment of 0.1, S_{op} from 0.4 to 0.6 at an increment of 0.1, and S_{cp} from 0.4 to 0.6 at an increment of 0.1. Once the first iteration returned the codebook entry with $OQ = OQ_s$ and $\alpha = \alpha_s$, the second iteration searched part of the larger codebook with an OQ value within $[OQ_s - 0.1, OQ_s + 0.1]$ and an α value within $[\alpha_s - 0.05, \alpha_s + 0.05]$.

For each audio recording, the first 50 ms segment (corresponding to the first 150 frames of high-speed recording) was processed and the F_0 was extracted using the Straight algorithm [22] with 25 ms window size and 1 ms window shift. The F_0 was then averaged for the first 50 ms segment. The harmonic magnitudes were calculated based on the averaged F_0. A Hamming window consisting of 4 pitch periods was used to calculate the spectrum of the input signal. The harmonic magnitudes were calculated in the range of 0-2600 Hz. This range is associated with the number of harmonics that can be reliably estimated from the spectrum. The window step size was 10 ms and the source estimation procedure was performed for each window. The final source waveform and OQ were obtained by averaging across the estimated source waveforms and OQ values over the first 50 ms segment.

4. Results

For comparison, the software toolkit Aparat [1] was used as a reference to obtained inverse-filtered source signals using Iterative and Adaptive Inverse Filtering (IAIF) [23]. Parameters were manually adjusted to minimize ripples in the inverse-filtered time waveform. The first 50 ms of each audio recording were inverse filtered. The cycle boundaries of the resultant glottal flow signal were marked and an average waveform was obtained by averaging across the cycles.

Our previous source estimation approach [2] where formant frequencies were estimated from the Snack toolkit [24] was also used for comparison. The source model and the codebook in that study are also updated as described in Section 3.2.

Table 1 shows the source estimation results in terms of Mean Square Error (MSE) between the estimated source waveform and the reference glottal flow waveform. Three estimation methods are shown: “Proposed” denoted the the method proposed in this paper, “Previous” denotes the estimation method in [2], and “Aparat” denotes manual inverse filtering using Aparat. Note that all the source waveforms are normalized both in time and amplitude for MSE calculation. Each waveform is 1000 samples in time with a maximum amplitude of 1 and a minimum amplitude of 0.

Under clean and 3 noise levels, the averaged MSE is lower for the proposed estimation algorithm than those of the previous approach and Aparat, on average. In clean condition, a statistical analysis shows that the differences in MSE among the three methods are not significant ($p > 0.01$). The performance improvement comparing the proposed approach to the previous

Table 1: *Results of the source estimation in terms of waveform MSE averaged across all the recordings (in %).*

	clean	20 dB	10 dB	5 dB
Proposed	6.6	6.8	8.4	9.6
Previous	6.9	7.9	10.2	11.7
Aparat	8.8	9.2	11.7	13.7

Table 2: *Results of the source estimation in terms of waveform MSE for each phonation type and pitch level (in %). B, M, and P denote three phonation types: breathy, modal, and pressed. L, N, and H denote three pitch levels: low, normal, and high.*

		phonation type			pitch level		
		B	M	P	L	N	H
clean	Proposed	5.7	6.8	7.5	6.8	6.3	6.8
	Previous	3.1	8.1	10.2	5.1	5.5	9.9
	Aparat	5.8	8.5	13.1	7.8	9.9	8.5
5 dB	Proposed	8.8	11.5	9.1	7.7	8.9	12.0
	Previous	8.4	13.1	14.3	10.7	12.1	12.2
	Aparat	12.5	13.9	15.1	14.1	13.1	14.0

approach increases with SNR level, partially due to the inaccuracy of formant estimation under noise. Under 5dB SNR, a statistically significant ($p < 0.01$) MSE improvement of 4.1% is observed when comparing the proposed approach to Aparat, suggesting the proposed algorithm is robust under white noise conditions. The performance of the proposed approach is also significantly better than that of the previous approach by 2.1% ($p < 0.01$).

Table 2 shows the MSE averaged within each phonation type and pitch level in clean and 5 dB SNR conditions. The MSE of the previous approach is 3.1% higher than that of the proposed approach for high-pitched cases (H) in clean condition ($p < 0.01$), highlighting the inaccuracies of LPC-based formant estimators for high-pitched voices. In clean condition, the proposed approach has higher MSE than the previous approach for three categories: breathy phonation (B), low pitch (L), and normal pitch (N), but the effect is not significant ($p > 0.01$). In 5 dB SNR condition, the proposed algorithm has lower MSE than Aparat for each phonation type and pitch level, and the effects are significant ($p < 0.05$) with the exception of modal phonation (M) and high-pitched cases (H).

Table 3: *The OQ estimation error for each phonation type and pitch level, and for each gender. B, M, and P denote three phonation types: breathy, modal, and pressed. L, N, and H denote three pitch levels: low, normal, and high.*

		phonation type			pitch level		
		B	M	P	L	N	H
clean	Male	.035	.072	.107	.025	.053	.082
	Female	.083	.049	.155	.045	.098	.148
5 dB	Male	.064	.092	.120	.035	.063	.084
	Female	.092	.108	.207	.104	.123	.161

Table 3 shows the OQ estimation error for each phonation type and pitch level, and for each gender. In both clean and 5 dB SNR conditions, the highest OQ estimation error occurs for high-pitched cases (H) and pressed phonations (P). On average, the OQ estimation error is higher for females than males. Recall that OQ ranges from 0 to 1.

5. Conclusions

This paper presents a new glottal flow model and a noise-robust source estimation method inspired by our earlier study [2]. The source signal is estimated using a codebook search approach. The glottal area extracted from high-speed images was converted to glottal flow to calibrate the proposed algorithm. Results in both clean and noisy conditions, on average, show that the proposed algorithm is robust in accurately estimating the glottal flow waveform. This study also provides an approach to explore the speech production chain by linking glottal area, glottal flow, and the acoustic speech signal.

Time-domain source models lack an effective way of modeling the incomplete glottal closure phenomenon, which has been shown to be an important physiological cue of voice production [25, 26]. Future work will include incorporating the incomplete glottal closure effect into the source modeling, as well as examing the effect of other noise types, for example babble noise.

6. References

[1] M. Airas, "TKK Aparat: An environment for voice inverse filtering and parameterization," *Logopedics Phoniatrics Vocology*, vol. 33, pp. 49–64, 2008.

[2] Y.-L. Shue and A. Alwan, "A new voice source model based on high-speed imaging and its application to voice source estimation," in *ICASSP*, 2010, pp. 5134–5137.

[3] A. Rosenberg, "Effects of the glottal pulse shape on the quality of natural vowels," *J. Acoust. Soc. Am.*, vol. 49, pp. 583–590, 1971.

[4] G. Fant, J. Liljencrants, and Q. Lin, "A four-parameter model of glottal flow," *STL-QPSR*, vol. 4, pp. 1–13, 1985.

[5] H. Fujisaki and M. Ljungqvist, "Proposal and evaluation of models for the glottal source waveform," in *ICASSP*, 1986, pp. 1605–1608.

[6] Y.-L. Shue, J. Kreiman, and A. Alwan, "A novel codebook search technique for estimating the open quotient," in *Interspeech*, 2009, pp. 2895–2898.

[7] G. Fant, *Acoustic theory of speech production*, Mouton, The Hague, Paris, 2nd edition, 1970, pp. 15-20.

[8] P. Alku, "Parameterisation methods of the glottal flow estimated by inverse filtering," in *VOQUAL*, 2003, pp. 81–87.

[9] J. Kane, M. Kane, and C. Gobl, "A spectral LF model based approach to voice source parameterisation," in *Interspeech*, 2010, pp. 2606–2609.

[10] E. Moore and M. Clements, "Algorithm for automatic glottal waveform estimation without the reliance on precise glottal closure information," in *ICASSP*, 2004, pp. 101–104.

[11] 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. Am.*, vol. 110, pp. 479–488, 2001.

[12] T. V. Ananthapadmanabha and G. Fant, "Calculation of true glottal flow and its components," *Speech Comm.*, vol. 1, pp. 167–184, 1982.

[13] I. Titze and B. Story, "Acoustic interactions of the voice source with the lower vocal tract," *J. Acoust. Soc. Am.*, vol. 101, pp. 2234–2243, 1997.

[14] K. N. Stevens, *Acoustic Phonetics*, MIT Press, Cambridge, MA, 1998, pp. 55–126.

[15] M. Rothenberg, *Acoustic interaction between the glottal source and the vocal tract, Vocal fold physiology*, University of Tokyo Press, Tokyo, 1981, pp. 305–323.

[16] B. Story and I. Titze, "Voice simulation with a body cover model of the vocal folds," *J. Acoust. Soc. Am.*, vol. 97, pp. 1249–1260, 1995.

[17] I. Titze and B. Story, "Rules for controlling low-dimensional vocal fold models with muscle activation," *J. Acoust. Soc. Am.*, vol. 112, pp. 1064–1076, 2002.

[18] Y.-L. Shue, *The Voice Source in Speech Production: Data, Analysis and Models*, Ph.D. thesis, University of California Los Angeles, 2010.

[19] B. Story, "Letalker: Matlab code for three-mass lumped-element model," 2011, `http://sal.shs.arizona.edu/~bstory/LeTalkerMain.html` (last viewed Apr. 2012).

[20] B. Story, I. Titze, and E. Hoffman, "Vocal tract area functions from magnetic resonance imaging," *J. Acoust. Soc. Am.*, vol. 100, pp. 537–554, 1996.

[21] J.W. Hawks and J.D. Miller, "A formant bandwidth estimation procedure for vowel synthesis," *J. Acoust. Soc. Am.*, vol. 97, pp. 1343–1344, 1995.

[22] H. Kawahara, A. de Cheveign, and R.D. Patterson, "An instantaneous–frequency–based pitch extraction method for high quality speech transformation: revised TEMPO in the STRAIGHT–suite," in *ICSLP*, 1998, paper 0659.

[23] P. Alku, "Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering," *Speech Comm.*, vol. 11, pp. 109–118, 1992.

[24] K. Sjölander, "The snack sound toolkit," KTH Stockholm, Sweden, 2004, `http://www.speech.kth.se/snack/` (last viewed Aug. 2009).

[25] G. Chen, J. Kreiman, Y.-L. Shue, and A. Alwan, "Acoustic correlates of glottal gaps," in *Interspeech*, 2011, pp. 2673–2676.

[26] B. Cranen and J. Schroeter, "Modeling a leaky glottis," *J. Phonetics*, vol. 23, pp. 165–177, 1995.