Proceedings of Meetings on Acoustics

Volume 19, 2013 http://acousticalsociety.org/

ICA 2013 Montreal
Montreal, Canada
2 - 7 June 2013

Speech Communication
Session 4aSCb: Voice and F0 Across Tasks (Poster Session)

4aSCb7. A perceptually and physiologically motivated voice source model

Gang Chen*, Marc Garellek, Jody Kreiman, Bruce R. Gerratt and Abeer Alwan

***Corresponding author's address: Electrical Engineering, Univ of California, Los Angeles, Los Angeles, CA 90095, gangchen@ee.ucla.edu**

Many glottal source models have been proposed, but none has been systematically validated perceptually. Our previous work showed that model fitting of the negative peak of the flow derivative is the most important predictor of perceptual similarity to the target voice. In this study, a new voice source model is proposed to capture perceptually-important source shape aspects. This new model, along with four other source models, was fitted to 40 voice sources (20 male) obtained by inverse filtering and analysis-by-synthesis (AbS) of samples of natural normal and pathologic phonation. We generated synthetic copies of the voices using each modeled source pulse, with all other synthesis parameters held constant, and then conducted a visual sort-and-rate task in which listeners assessed the extent of perceived similarity between the target voice samples and each copy. Results showed that the proposed model provided a more accurate fit and a better perceptual match to the target than did the other models.

Published by the Acoustical Society of America through the American Institute of Physics

©2013 Acoustical Society of America [DOI: 10.1121/1.4798784]
Received 18 Jan 2013; published 2 Jun 2013

INTRODUCTION

According to the linear speech production model [1], speech signals are generated by filtering the voice source by the vocal tract transfer function. Modeling the glottal source has been an important topic for decades and has applications in many areas, such as speech coding and speech synthesis. Many source models have been proposed with varying levels of complexity, such as the Rosenberg [2], Liljencrants-Fant (LF) [3], and Fujisaki-Ljungqvist (FL) [4] models (see [5] for review). With three parameters, the Rosenberg trigonometric model (denoted Ros) has two separate functions for the opening and closing phases to represent the glottal flow volume velocity [2]. The LF and FL models represent the first derivative of the glottal volume velocity pulse, which incorporates lip radiation effects. The simplified four-parameter LF model [3] uses a combination of sinusoidal and exponential functions, and is commonly used in speech synthesis. With six parameters and polynomial functions, the FL model provides greater detail in modeling the glottal pulse shape, but the increased number of parameters also makes it more difficult to use in practice. The four-parameter glottal flow model (denoted EE1 [6]) uses a combination of sinusoidal and exponential functions similar to the LF model, but with the ability to adjust the slopes of the opening and closing phases separately. The glottal flow model in [7] (denoted EE2) improves the EE1 model by redefining the model parameters (speed of opening and speed of closing) to allow for lower computational complexity, faster waveform generation, and more accurate pulse shape manipulation. In that study, the EE2 model was used for automatic glottal flow estimation from acoustic speech signals, and glottal area waveforms extracted from high-speed endoscopic recordings of the laryngeal vibrations were converted to glottal flow in order to evaluate the performance of the glottal flow estimation algorithm.

Research efforts have also been devoted to studying the perceptual importance of changes in source waveform shapes. In [2], listening tests using a variety of glottal excitations showed that simulated excitations with a single slope discontinuity at closure were perceived as more natural-sounding, while very small opening or closing times (or opening times approximately equal to or less than closing times) were not preferred. These experiments also demonstrated that, not surprisingly, glottal models that closely resembled the characteristics of the natural glottal pulse resulted in better synthetic speech. In [8], the LF model and a turbulent noise generator were used to synthesize four voice quality types (modal, vocal fry, falsetto, and breathy). Perceptual experiments showed that these four voice quality types could be characterized by four parameters: pulse width, pulse skewness, the abruptness of glottal closure, and turbulent noise.

Few studies have attempted to systematically validate glottal source models perceptually, and model development has focused more on replicating observed pulse shapes than on perceptual sufficiency. As a result, it is unclear which (if any) deviations from perfect fit between models and data have perceptual importance. In our previous study [9], the Ros, FL, LF, EE1, and EE2 source models were fitted to 40 natural normal and pathological voice sources (20 male) obtained by inverse filtering and analysis-by-synthesis (AbS), subject to mean square error (MSE) criteria for which each point of the waveform was weighted equally. Evaluation of model fit at different parts of the source waveforms showed that the fit to the target pulses was worst at the negative peak of the flow derivative. Synthetic copies of the voices were then created using each modeled source pulse, while holding all other synthesizer parameters constant (including formant frequencies and bandwidths, fundamental frequency (F0) and amplitude contours, and spectral noise levels). These stimuli were compared to the AbS target in a sort-and-rate listening test (described below). Across models and voices, the perceptual match between the target and synthetic tokens was best predicted by the match between the target and modeled stimuli at the negative peak of the flow derivative ($R^2 = 0.34$). Fit during the opening phase also contributed weakly but significantly ($p < 0.01$) to the perceptual match.

In a follow-up experiment, we fitted the models to the AbS sources subject to MSE criteria while constraining the models to fit the negative peak of the flow derivative precisely, which significantly increased the mismatch to the opening phase ($p < 0.01$; see Figure 1). Informal listening tests on several tokens showed that this significant mismatch to the opening phase resulted in a noticeable perceptual difference between the target and modeled stimuli. These results indicate the need for a source model with increased flexibility to provide a close fit to all parts of the voice source signal.

In this study, a new voice source model, motivated by data from high-speed laryngeal videoendoscopy, is proposed to capture perceptually-important source shape aspects. This model is then evaluated in comparison to 4 existing source models, with respect to fit in both the MSE and perceptual senses. Finally, an automated approach to model fitting was evaluated with respect to a set of voice samples for which quality and F0 varied orthogonally, to assess the applicability of the proposed model in automatic applications.

DATA AND METHOD

Stimuli

Source model comparisons required a target source pulse to which the models could be fitted, and the need for experimental control during perceptual evaluations mandated that this target be synthetic, so that voice stimuli could be created that differed in the source, with all other parameters held constant. To ensure that these synthetic targets were as natural in quality as possible and that they represented a range of naturally-occurring voice qualities, target stimuli were derived via analysis-by-synthesis (AbS [10]) from 40 natural samples (20 M) of the vowel /a/. Samples were directly digitized at 20 kHz using a Brüel & Kjær microphone (model 4193), and a 1-second-long segment was excerpted for analysis. Briefly, the synthesizer sampling rate was fixed at 10 kHz. Parameters describing the harmonic part of the voice source were estimated by inverse filtering a representative cycle of phonation for each voice using the method described in [11]. The spectral characteristics of the inharmonic part of the source (the noise excitation) were estimated using cepstral-domain analysis similar to that described in [12]. Spectrally-shaped noise was synthesized by passing white noise through a 100-tap finite impulse response filter fitted to that noise spectrum. To model the F0 and amplitude contours, F0 was tracked pulse by pulse on the time domain waveform. Formant frequencies and bandwidths were estimated using autocorrelation linear predictive coding analysis with a window of 25.6 ms. The complete synthesized source was then filtered through the vocal tract model, and all parameters were adjusted until the synthetic copy formed an acceptable match to the original natural voice sample. A paired comparison (same/different) task ensured that the AbS tokens were indistinguishable from the natural stimuli: d prime ranged from 0 to 1.32 across voices, with a mean of 0.79 (sd=0.41). Given these results, the AbS tokens were used in place of the natural voice samples as the target stimuli in all subsequent analyses.

The Proposed Model

The proposed model is based on the models in [6, 7], which were motivated by shapes of glottal area waveforms extracted from laryngeal high-speed videoendoscopy. The model is a combination of sinusoidal and exponential functions shown to be effective in approximating a wide range of glottal flow pulse shapes. The model is then refined using AbS to eventually capture the shapes of the glottal flow derivative, as the LF model does. The model has six parameters: the time of the positive peak (t_i), the shape of the opening (S_1; amplitude of the waveform at $t_i/2$), the time of the peak flow (t_p; zero-crossing of the flow derivative), the time of the negative peak (t_e), the amplitude of the negative peak (E_e), and the slope of the return

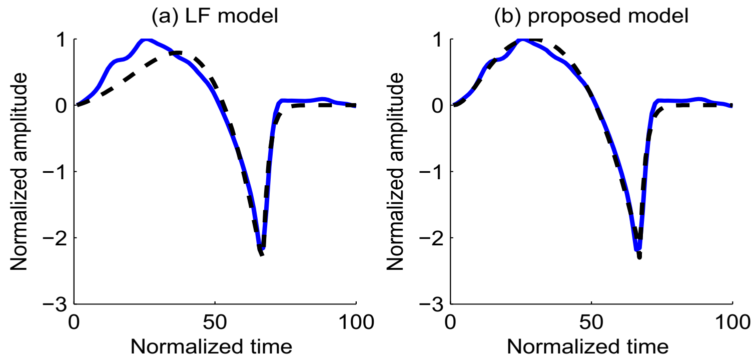

FIGURE 1: An example of fitting the LF and the proposed models to the same AbS source pulse subject to MSE criteria while constraining the models to fit the negative peak of the flow derivative precisely. Solid line: AbS source. Dashed line: model-fitted source.

phase (t_a). The latter four parameters (t_p, t_e, E_e, and t_a) were originally defined in the four-parameter LF model [3]. The first two parameters were added to the proposed model to provide an additional degree of freedom, so that the timing of the positive peak and the shape from the start to the positive peak can be manipulated directly, independent of the negative peak of the flow derivative. With these parameters, the glottal opening phase could be modeled more accurately. Recall that our previous studies showed that a significant mismatch to the opening phase could lead to a noticeable perceptual difference between the target and the modeled stimuli. An example of a model waveform is shown in Figure 2. Given the six parameters described above, mathematically the glottal flow derivative $u(t)$ is defined as:

$$u(t) = \begin{cases} f(\frac{t}{t_i}, \lambda_1) & (0 \le t \le t_i) \\ [f(\frac{(2t_e - t_i - t)}{2(t_e - t_i)}, \lambda_2) - 1]\frac{12(1+E_e)}{6+\lambda_2} + 1 & (t_i < t \le t_e) \\ \frac{-E_e}{\epsilon t_a}[e^{-\epsilon(t-t_e)} - e^{-\epsilon(t_c - t_e)}] & (t_e < t \le 1) \end{cases}$$

$$\begin{aligned} f(t,\lambda) &= \frac{1}{\pi(e^{\lambda}+1)}\{e^{\lambda t}[\lambda sin(\pi t) - \pi cos(\pi t)] + \pi\} \\ \lambda_1 &= 12 \cdot (0.5 - S_1) \\ \lambda_2 &= \arg\min_{\lambda} \left| f(\frac{2t_e - t_p - t_i}{2(t_e - t_i)}, \lambda) - \frac{12E_e + 6 - \lambda}{12(E_e+1)} \right| \\ \epsilon &= \frac{1}{t_a}[1 - e^{-(t_c - t_e)/t_a}] \end{aligned}$$

t_c is the time of closure. In practice it is convenient to set $t_c = 1$, i.e., the complete fundamental period [3]. ϵ, λ_1 and λ_2 are intermediate parameters. As illustrated in Figure 2, the proposed parameters can be easily derived from the inverse-filtered differential glottal waveform, and directly control the shape of the glottal waveform in a straightforward way. Unlike the LF model, which describes the open phase ($0 < t < t_e$) using one function, the proposed model uses two functions ($0 < t < t_i$ and $t_i < t < t_e$) to describe the open phase, allowing for more flexibility in modeling. Figure 1 (b) shows an example of constraining the proposed model to fit the negative peak of the flow derivative precisely, while still achieving satisfactory fittings in other parts.

Model Fitting

In this study, each of the 40 target AbS-derived source functions was fitted with 5 source models: the Ros, LF, EE1, EE2, and the proposed model. The FL model, which provided the

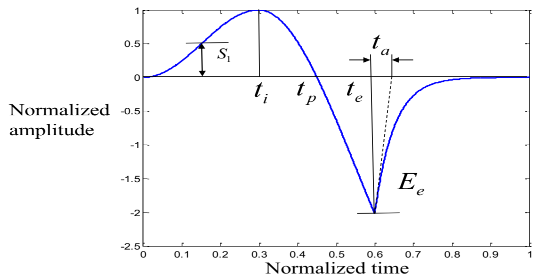

FIGURE 2: An example of the proposed model with $S_1 = 0.5, t_i = 0.3, t_p = 0.45, t_e = 0.6, E_e = 2$, and $t_a = 0.05$.

worst fit to the target sources in our previous experiment, was excluded from further experiments. First-derivative representations were calculated mathematically for the Ros, EE1, and EE2 models, which describe flow pulses in the time domain, so that all models were fitted to the target AbS source functions in the flow derivative domain. One cycle of the AbS source signal for each speaker was normalized to a maximum amplitude of 1. Each derivative-domain model was fitted to all of the AbS source functions using MSE criteria, for which each point of the waveform was weighted equally. Additionally, the proposed model was fitted a second time to the AbS source function with the constraint of exactly matching the first point, the positive peak of the flow derivative, the time of maximum flow (zero-crossing of flow derivative), and the negative peak of the flow derivative. Note that it is not always possible to exactly match ALL landmarks for the other models, due to constraints inherent in the models and their parameters. Because of the increased flexibility, especially in modeling the opening phase, the proposed model is able to match all landmarks well. This condition was included for the proposed model in order to assess the combined perceptual importance of the landmarks of the voice source signal. Target AbS source pulses and the corresponding least-MSE-fitted sources using the proposed model for six different speakers are shown in Figure 3. As this figure shows, the proposed model is able to approximate a wide range of pulse widths, pulse skewnesses, and abruptnesses of glottal closure.

Perceptual Experiment

To determine the perceptual importance of these results, we generated synthetic copies of the voices using each modeled source pulse for each voice, with all other synthesizer parameters held constant at the values derived during AbS, as illustrated in Figure 4. For the proposed model, only the model-fitted sources with exact matching at the landmark points were used in this experiment (denoted "Proposed-LM"). 40 listeners (UCLA students and staff; 18-33 years of age; M=21.15 years; sd=3.03 years) assessed the similarity of all versions of each voice in a visual sort-and-rate task [13, 14], in which listeners assessed the extent of perceived match between the original voice samples and each copy. Each listener heard 10 voice "families", where each family included an original natural voice sample, the corresponding target AbS token, and the 5 model-synthesized tokens of the same voice, such that across subjects each family was judged by 10 listeners. The stimuli were presented as distinct icons on the screen. For each family (each trial), listeners were asked to play the stimuli by clicking the icons, and to place perceptually similar sounds close together on a line on the screen, while perceptually dissimilar sounds were to be placed farther away. Listeners were instructed to use as much of the line for sorting the stimuli as they wished. They could listen to the stimuli as often as they like, and the

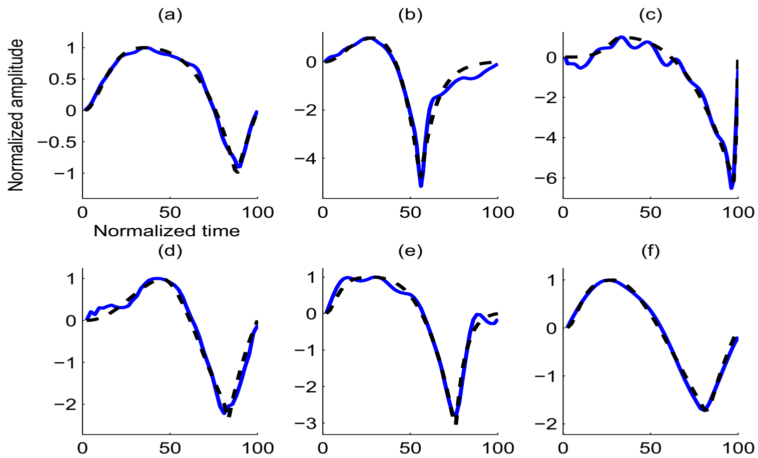

FIGURE 3: Target AbS source pulses and the corresponding least-MSE-fitted sources using the proposed model for six different speakers. Panels (a), (b), and (c): male speakers. Panels (d), (e), and (f): female speakers. Solid line: AbS source. Dashed line: the proposed model.

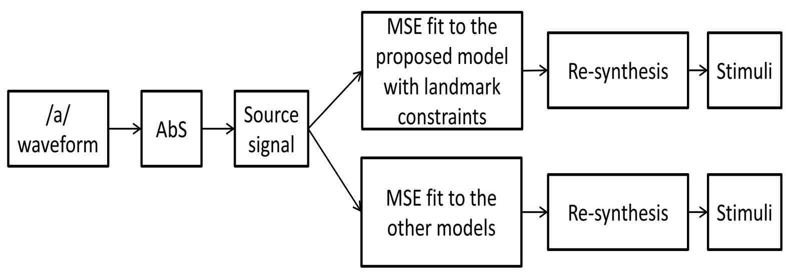

FIGURE 4: Flowchart showing how stimuli were generated for the perceptual experiment.

study was not timed.

Although listeners saw no numerical values associated with the endpoints of the line, the left and right endpoints were assigned values of 0 and 1000, respectively. Thus, a numerical value could be assigned to the position of each token. We then calculated the distance of each modeled token from the target AbS voice, and this value was subsequently normalized within family for the range of values used on that given trial by that listener. The absolute values of these normalized distances were used in subsequent analyses, because orientation of the line was arbitrary and varied from listener to listener.

RESULTS

Overall Model Fit

Table 1 shows MSE values for fit of each of the source models under study to the target AbS sources. (See table caption for the meaning of model labels.) A two-way repeated measures ANOVA (model by speaker sex) showed significant main effects of model $[F(5,190) = 12.99, p < 0.0001]$ and sex $[F(1,38) = 8.71, p < 0.01]$ on mean MSE, as well as a significant model by sex interaction effect $[F(5,190) = 4.27, p < 0.001]$. Tukey post-hoc t-tests (with Bonferroni adjustment for multiple comparisons) indicated that no cross-model differences were significant for female speakers. For male speakers, a separate t-test showed that the "Proposed" model had lower MSE values than the Ros, LF, EE1, and EE2 models ($p < 0.05$).

TABLE 1: MSE values (in %) of fitting models to the AbS sources. "Proposed" denotes fitting the proposed model subject to MSE criteria. "Proposed-LM" denotes fitting the proposed model subject to MSE criteria with the constraint of exact landmark matching.

	Ros	LF	EE1	EE2	Proposed	Proposed-LM
Male	27.8	14.1	25.8	21.6	3.9	6.9
Female	11.3	3.6	3.8	3.5	1.2	1.6

Perceptual Experiment

Results of the perceptual experiment are shown in Table 2. Recall that 40 listeners participated in this task, but each only heard 10 of the 40 voices. Thus, every 4 subjects heard the stimuli from all 40 voices. Because a pre-test showed no significant differences in rating, we averaged the results of every 4 subjects, to make 10 "metasubjects", where each "metasubject" (consisting of 4 listeners) heard all 40 voices. This enabled us to run an ANOVA with "metasubject" as the error term. A two-way (model by sex of voice) repeated-measures ANOVA showed significant main effects of model [$F(4,36) = 155.77, p < 0.0001$] and sex [$F(1,9) = 26.49, p < 0.001$] on mean perceptual distance, as well as a significant model by sex interaction effect [$F(4,36) = 10.62, p < 0.001$]. Tukey post-hoc t-tests (with Bonferroni adjustment for multiple comparisons) indicated that the proposed-LM model formed a significantly better match to the target AbS stimulus (lower mean perceptual distance) than the other models ($p < 0.0001$). The difference between male and female voices in perceptual distances between the modeled and target tokens was significant only for the Ros model, for which male voices were closer perceptual matches to the AbS voice than female voices ($p < 0.0001$). For both sexes, the Ros model had a higher perceptual distance than the other models ($p < 0.0001$).

TABLE 2: Normalized perceptual distances (range from 0 to 1) between the model-fitted voices and the target AbS voice, for male and female voices. A smaller number indicates a closer perceptual distance (closer match) to the target AbS voice.

	Ros	LF	EE1	EE2	Proposed-LM
Male	0.57	0.46	0.38	0.40	0.26
Female	0.71	0.42	0.46	0.43	0.32

Model Fitting to Sources with Different Voice Qualities

Gathering perceptually-corrected AbS data requires substantial manual effort, and is often impractical in voice source modeling or synthesis applications. In this section, we explore model fit to "raw" or "noisy" data, using automatic analyses without time-consuming manual intervention. Further, these analyses examined utterances for which voice quality was systematically varied by experienced speakers to represent continua from breathy to pressed and from low F0 to high F0. Although voice quality is an aspect of perceived naturalness, most source models were originally developed to represent voices with modal (normal) quality. By assessing model fit across a balanced set of quality and F0 variations, we hope to demonstrate the potential applicability of this model for improving the quality of speech synthesis

The speech data were the same as those in [15]. Briefly, 6 subjects (3 male/3 female) were asked to vary their F0 (low, normal and high) and voice quality (pressed, normal and breathy) quasi-orthogonally while sustaining the vowel /i/, resulting in 9 tokens from each speaker. The most stable 1 second of phonation was extracted for analysis. The source signals were obtained using the automatic inverse-filtering software toolkit Aparat [16]. Parameters were manually adjusted to minimize ripples in the inverse-filtered time waveform; no AbS was conducted to perceptually correct the source signals. The LF model and the proposed model were then fitted to source signals representing different voice qualities and F0 levels, as described above, subject to MSE criteria.

Table 3 shows MSE values of fitting the LF and proposed models to the source signals with different voice qualities and F0 levels. Although the six subjects were directed to produce a total of 9 different phonations, one female was not able to produce a pressed phonation with normal F0, and one male was unable to produce a low F0 for the three voice qualities. We therefore did not run an ANOVA for this section, because the data were not balanced. T-tests (corrected for multiple comparisons) showed that the proposed model had a significantly lower MSE value than the LF model for high-F0 voices ($p < 0.05$) and breathy voices ($p < 0.05$). The MSE values did not differ significantly between the two models for the other F0 levels and voice qualities. The fitting to breathy voices had significantly higher ($p < 0.01$) MSE values than normal and pressed voices for both models. One possible explanation for these findings is that the aspiration noise generated with breathy voices might have affected the accuracy of inverse filtering. Another reason might be the smoothed glottal closures in breathy voices, because most source models assume a slope discontinuity of the first time derivative (e.g., point $(t_e, -E_e)$ of Figure 2). Although it was reported in [2] that one slope discontinuity at closure was preferred to generate good-quality synthetic speech (normal voice quality), it is not clear whether this is essential for synthesizing breathy voice, where the closure is more gradual and smooth. Perceptual experiments using breathy voice generated from source signals with different pulse shapes at closure would aid further investigation, and could help determine whether the differences in model fitting across voice qualities and F0 levels are perceptually significant.

TABLE 3: MSE values (in %) of fitting models to the source signals with different voice qualities and F0 levels.

	LF	Proposed		LF	Proposed
Breathy	16.8	11.2	High F0	12.9	7.8
Modal	12.4	8.7	Normal F0	13.8	10.4
Pressed	8.6	5.6	Low F0	10.1	7.5

RELATION TO PRIOR WORK

This paper presented a systematic perceptual evaluation of various source models, and proposed a new model to capture perceptually-relevant information. The study in [8] investigated the factors of vocal quality that might be affected by changes in voice source signals but only 3 listeners were involved. In that study, only the LF model was used to generate the source signal. In [4], 6 models were evaluated but were only used in a task to minimize the linear predictive error from the original voice. In this study, 5 models were evaluated in terms of both physical fits (MSE) to the AbS source and perceptual matches to the target AbS stimuli. Results were based on perceptual experiments with 40 listeners and 40 voice samples.

CONCLUSION AND FUTURE WORK

This study presented a new voice source model with increased flexibility to capture the perceptually-important source shape aspects. Five voice source models were fitted to 40 natural voices obtained by inverse filtering and analysis-by-synthesis (AbS). Synthetic copies of the voices were generated using each modeled source pulse. Models were perceptually evaluated using a visual sort-and-rate task in which listeners assessed the extent of perceived match between the AbS copies and stimuli created with model-fitted sources. Compared to the other models, on average, the proposed model provided more accurate fittings to the AbS-derived source, and source signals with three voice qualities (breathy, modal, and pressed). Perceptual experiments showed that the proposed model provided closer perceptual matches to the target AbS voice than the other models. Future work will examine the effect of using this model in synthesizing continuous speech.

ACKNOWLEDGEMENTS

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

REFERENCES

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

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

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

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

[5] K.E. Cummings and M.A. Clements, "Glottal models for digital speech processing: A historical survey and new results," *Digital Signal Processing*, vol. 5, pp. 21–42, 1995.

[6] 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.

[7] G. Chen, Y.-L. Shue, J. Kreiman, and A. Alwan, "Estimating the voice source in noise," in *Interspeech*, 2012,

[8] D.G. Childers and C.K. Lee, "Vocal quality factors: Analysis, synthesis, and perception," *J. Acoust. Soc. Am*, vol. 90, pp. 2394–2410, 1991.

[9] J. Kreiman, B.R. Gerratt, G. Chen, M. Garellek, and A. Alwan, "Perceptual evaluation of source models," *J. Acoust. Soc. Am*, vol. 132, p. 2088, 2012.

[10] J. Kreiman, N. Antoñanzas-Barroso, and B.R. Gerratt, "Integrated software for analysis and synthesis of voice quality," *Behavior Research Methods*, vol. 42, pp. 1030–1041, 2010.

[11] H.R. Javkin, N. Antoñanzas-Barroso, and I. Maddieson, "Digital inverse filtering for linguistic research," *J. Speech Hear. Res.*, vol. 30, pp. 122–129, 1987.

[12] G. de Krom, "A cepstrum-based technique for determining a harmonics-to-noise ratio in speech signals," *J. Speech Hear. Res.*, vol. 36, pp. 254–266, 1993.

[13] S. Granqvist, "The visual sort and rate method for perceptual evaluation in listening tests," *Logopedics Phonatrics Vocology*, vol. 28, pp. 109–116, 2003.

[14] C.M. Esposito, "The effects of linguistic experience on the perception of phonation," *J. Phonetics*, vol. 38, pp. 306–316, 2010.

[15] J. Kreiman, Y.-L. Shue, G. Chen, M. Iseli, B.R. Gerratt, J. Neubauer, and A. Alwan, "Variability in the relationships among voice quality, harmonic amplitudes, open quotient, and glottal area waveform shape in sustained phonation," *J. Acoust. Soc. Am.*, vol. 132, pp. 2625–2632, 2012.

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