A NEW VOICE SOURCE MODEL BASED ON HIGH-SPEED IMAGING AND ITS APPLICATION TO VOICE SOURCE ESTIMATION

Yen-Liang Shue and Abeer Alwan

Department of Electrical Engineering, University of California Los Angeles
405 Hilgard Ave., Los Angeles, CA 90095

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

ABSTRACT

There are numerous models of varying complexities which seek to efficiently represent the voice source signal. These models are typically based on data and observations which can come from air-flow masks, electroglottographs, mechanical systems, and the inverse-filtering of speech signals. The first part of this study examines observations from the high-speed imaging of the larynx and proposes a new source model, which is shown to provide a better fit for the observed data than existing models. The proposed source model is then used in an automatic source estimation application, based on methods introduced in an earlier study [1]. Results, on average, show that the proposed model provides a more accurate estimation of the source signal compared with the Liljencrants-Fant model.

Index Terms— source estimation, voice source, speech analysis

1. INTRODUCTION

The voice source signal is an essential part of the speech production process, containing a vast amount of non-lexical information. This information can convey, for example, prosodic events, emotional status, as well as cues pertaining to the uniqueness of the speaker's voice. In medical applications, analysis of the voice source can be used to diagnose diseases of the vocal cords. Generally, the acoustic source signal (i.e. glottal flow) is thought to be the result of non-linear interactions between the lung pressure and the glottal area function [2]. However, a recent study [3] found that the differences between the acoustic source pulse shapes and the glottal area waveforms were small relative to the larger differences across the area waveforms.

To effectively study the properties of the voice source, accurate source models are needed. Many source models with varying complexities have been proposed, such as the Rosenberg [4], Liljencrants-Fant (LF) [5], and the Fujisaki-Ljungqvist [6] models. A more detailed comparison between these and several other source models can be found in [7]. The motivations for such a wide range of models are mainly due to the different types of data and observations on which the models are built upon. These observations can come from air-flow masks, electroglottographs (EGG), mechanical systems, and inverse-filtering of speech signals based on the linear speech production model [8]. The first part of this study presents a new source model which is derived using glottal area waveforms from the high-speed imaging of the larynx.

The second part of this study applies the proposed source model to the automatic estimation of voice source waveforms from acoustic speech signals. Estimation of the source signal is a non-trivial task and is typically obtained using two main methods. The first method, as used in [9] and [10], involves estimating the vocal tract transfer function (VTTF) and inverse-filtering the original speech signal to obtain a residual signal, which is then fitted to a source model. The second method relies on a joint-estimation approach ([11], [12], [13], and [14]) in which the source parameters and the VTTF are estimated together in a global approach. The basic assumption in both of these methods is that speech can be approximated by a linear process in which a source function is filtered by the VTTF to produce the desired output. However, it is well known that during speech production, source-tract coupling occurs which can result in non-linear effects. In the first method, these non-linearities usually appear in the residual signal, which is then used for source-model fitting. In the joint-estimation method, the non-linearities may be incorporated into both the source parameters and the VTTF. Calibration of the algorithms in these methods are often performed with analysis-by-synthesis results or EGG signals, which are an indirect observation of the glottis that has inherent difficulties with extracting the "true" source signal.

In our earlier study [1], a new method was proposed in which the LF source signal was used effectively in inverse-filtering, resulting in a residual signal which was used to fit to the parameters of the VTTF. By reversing the roles of the source model and the VTTF, it was hoped that the non-linear effects and other noises could be mapped onto the parameters of the VTTF, providing a more accurate source estimate. In that study, the estimation of the open quotient (OQ) from the acoustic signals were compared with the measured values of OQ from high-speed imaging of the larynx. Inconsistencies in some of the results suggested that some modifications to the LF-model may be required to accurately model the observed vibration patterns in the glottis. In this study, the proposed source model is used to estimate the source signal based on the method in [1]. Results are calibrated with observations from the high-speed imaging.

2. DATA

The data used in this study is similar to that used in [1]; a summary of the data-collection process is presented here. Audio and high-speed video (3000 frames/second) were recorded synchronously from subjects who were asked to produce the vowel /i/ with different voice qualities (pressed, normal and breathy) and different F_0 (low, normal and high). In addition to the original 4 subjects, 2 more subjects were recorded for a total of 6 subjects (3 females, denoted by FM1–3, and 3 males, denoted by M1–3). For each recording, one second samples of audio and video were retained from the most stable sections for analysis.

Image segmentation was performed on the first 150 frames of

Work supported in part by the NSF.

each high-speed video sample to extract the open glottal area. This process was done manually to ensure accuracy. Each cycle of glottal vibration was then marked by recording, where it existed, the first instances of glottal opening. In samples where there were no complete glottal closures, the minimum glottal opening points were recorded. These points allowed the measurements of the open glottal areas to be averaged across the glottal cycles and produce a waveform which is representative of the source signal for the 150 analyzed frames for that utterance. Hence, there were 9 representative glottal area waveforms (3 F_0 types each with 3 phonation types) per subject.

3. A NEW VOICE SOURCE MODEL

The results in [1] suggested that the LF-model may not accurately describe some glottal area waveforms. This is not surprising given that the LF-model was derived from a different set of data. Inspections of the extracted waveforms revealed two limitations of the model: (1) it was noticed that the opening phase can be shorter in duration than the closing phase and (2), both the opening and closing phases of the LF-model are slow relative to the observed data, which showed that the vocal folds can open and close very quickly.

The proposed model consists of 5 parameters: the fundamental period (T_0), open quotient (OQ), asymmetry coefficient (α), speed of opening phase (S_{op}) and speed of closing phase (S_{cp}). An example of a model waveform is shown in Fig. 1. Using the notation from this figure, $OQ = \frac{t_o+t_c}{T_0}$, $\alpha = \frac{t_o}{t_o+t_c}$, $S_{op} = \frac{t_{oh}}{t_o}$ and $S_{cp} = 1 - \frac{t_{ch}}{t_c}$, where t_{ch} and t_{oh} are at 50% of the amplitude. With the exception of T_0, the four other parameters all range from 0 to 1. This model is a time-domain glottal flow model (i.e. not the flow derivative) and utilizes a variant of the integrated first LF-model equation, shown in Eq. 1 with the original notation from [5], for both the opening and closing phases.

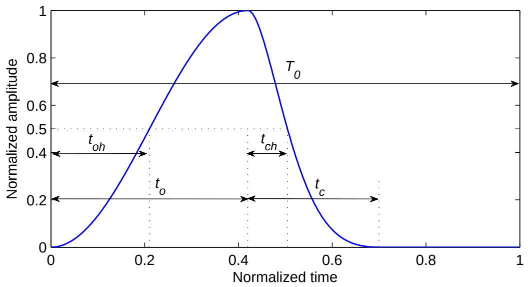

Fig. 1. Example of the proposed model with $OQ = 0.7$, $\alpha = 0.6$, $S_{op} = 0.5$ and $S_{cp} = 0.7$.

$$u_g(t) = \frac{E_0}{\alpha^2+\omega^2}\left[e^{\alpha t}(\alpha sin(\omega_g t) - \omega_g cos(\omega_g t)) + \omega_g\right] \quad (1)$$

The proposed model is defined as:

$$u(t) = \begin{cases} f(\beta_o t, \lambda_{S_{op}}), & 0 \le t \le \alpha OQ \cdot T_0 \\ f(\beta_c(OQ \cdot T_0 - t), \lambda_{S_{cp}}), & \alpha OQ \cdot T_0 < t \le OQ \cdot T_0 \\ 0, & OQ \cdot T_0 < t \le T_0 \end{cases}$$

where

$$f(x, \lambda^*) = A(\lambda^*)\left[e^{\lambda^* x}(\lambda^* sin(\pi x) - \pi cos(\pi x)) + \pi\right]$$

and

$$\lambda^* = \arg\min_{\lambda}\left|\frac{e^{\lambda s}(\lambda sin(\pi s) - \pi cos(\pi s))}{\pi(e^{\lambda}+1)} + \frac{1}{e^{\lambda}+1} - \frac{1}{2}\right|$$

with $A(\lambda^*) = (\pi(e^{\lambda^*}+1))^{-1}$, $s = S_{op}$ or S_{cp}, $\beta_o = (\alpha OQ \cdot T_0)^{-1}$, and $\beta_c = ((1-\alpha)OQ \cdot T_0)^{-1}$. λ^* determines the slope of $f(x, \lambda^*)$ which can be calculated by simple optimization techniques. A somewhat non-trivial closed-form solution for λ^* also exists involving the Lambert W-function. $A(\lambda^*)$ is a normalizing term so that $\max f(x, \lambda^*) = 1$.

It is important to note that while phonations with incomplete glottal closures (i.e. glottal gaps) could be modeled by adding a "DC-offset" parameter, it was not done here because: (1) radiation is often modeled by applying a derivative operation to the glottal flow signal which would remove the effects of the DC-offset, and (2) glottal gaps are perceived as turbulent noise, which is not modeled here.

3.1. Model fitting performance

The average glottal area waveforms obtained in Sec. 2 were first normalized to have a minimum value of 0 (i.e. DC-offset removed), a maximum value of 1, and were resampled to a length of 100 samples. For each subject, the proposed model was fitted to the normalized source pulses by using a mean squared error (MSE) criterion. For comparison purposes, the LF-model was also fitted to the normalized pulses for each subject. The average MSE across all speakers and utterances for the proposed model was 0.001 while for the LF-model, the average MSE was 0.011. As expected, visual inspections of the fitted model waveforms also showed that the proposed model provides a more accurate fit of the glottal area waveform in all cases. Fig. 2 shows two examples of the fitting results. Note the top panel shows a source waveform with an opening phase duration which is much smaller than the closing phase duration. This type of waveform was found consistently for all 6 subjects, especially during the normal and breathy phonations. These results show that while the LF-model does not fit the glottal area data well, it does provide a good basis for the derivation of newer models.

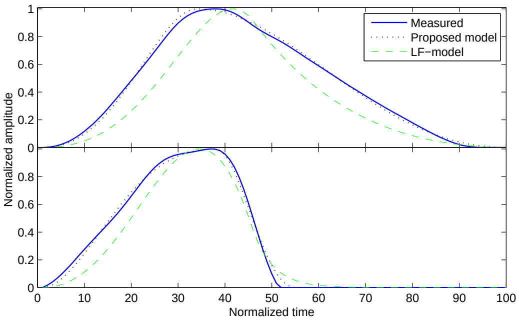

Fig. 2. Examples of fitting results for the proposed source model and the LF-model. The top panel is from the low F_0, normal phonation of subject FM1 and the bottom panel is from the high F_0, normal phonation of subject M3.

4. APPLICATION TO SOURCE ESTIMATION

In this section, the proposed source model is used to create a codebook for an automatic source signal estimation experiment.

4.1. Method

The method used for automatic source estimation is based on the harmonic magnitude matching technique described in [1]. The main block diagram is shown in Fig. 3. In this method, a codebook of source signals is used to implicitly inverse filter an input signal, leaving the residual signal for the VTTF and other non-linear source-tract interactions. Briefly, for an input signal, the harmonic magnitudes of its spectrum are calculated and normalized to the first harmonic magnitude; this is denoted by S_n (for the n–th normalized harmonic magnitude) in Fig. 3. A similarly normalized source signal from a given codebook, denoted by U_n^k for the k–th entry, is subtracted from S_n resulting in a residual signal, V_n, which is then used in a constrained optimization operation to find the vocal tract parameters; as with the earlier study, a 3–formant VTTF was used. The source which results in the lowest analysis-by-synthesis error is selected.

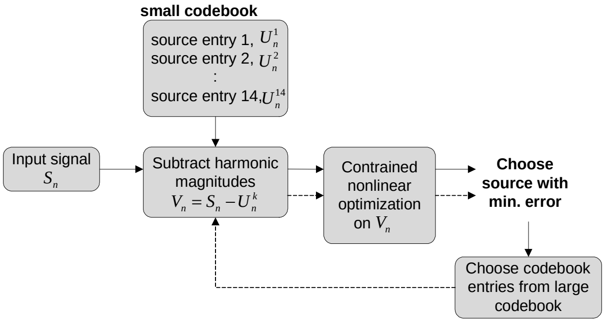
small codebook
source entry 1, U_n^1
source entry 2, U_n^2
:
source entry 14,U_n^{14}
Input signal S_n
Subtract harmonic magnitudes $V_n = S_n - U_n^k$
Contrained nonlinear optimization on V_n
Choose source with min. error
Choose codebook entries from large codebook

Fig. 3. Block diagram showing the two iterations of the method [1]; solid and dashed lines represent the first and second iteration, respectively.

As in [1], two iterations of the algorithm were used to reduce the overall processing time. In the first iteration of the algorithm, shown with solid lines in Fig. 3, a smaller codebook was used to find the approximate source parameters. The smaller codebook was generated by performing a grid search over the two parameters OQ and α at the following resolutions: OQ from 0.4 to 1.0 at increments of 0.1, and α from 0.4 to 0.5 at increments of 0.1. S_{op} and S_{cp} were both set to a constant value of 0.5. This resulted in a codebook of 14 entries. The codebook entry selected at the end of the first iteration was used to find source candidates from a larger codebook with finer parameter settings. Assuming that the OQ and α values of the selected source entry were denoted by OQ_s and α_s, sources in the larger codebook with an OQ value within $OQ_s \pm 0.1$ and an α value within $\alpha_s \pm 0.05$ were selected for the second iteration. This larger codebook was generated with the following parameter resolutions: OQ from 0.35 to 1.00 at increments of 0.01, α from 0.35 to 0.50 at increments of 0.05, S_{op} from 0.3 to 0.7 at increments of 0.2, and S_{cp} from 0.3 to 0.7 at increments of 0.2. This produced a codebook of size 2376 which was further reduced to 2179 entries by performing a correlation analysis and removing entries which had a cross-correlation coefficient of 0.999 or more. α was only given half of its possible range due to the property of the Fourier transform for time-reversed signals; e.g., for a signal $h(t)$ with periodicity T_0, $|\mathcal{F}\{h(t)\}| = |\mathcal{F}\{h(T_0 - t)\}|$, where $\mathcal{F}$ denotes the Fourier transform, and a source with $\alpha = 0.4$, $S_{op} = 0.6$ and $S_{cp} = 0.5$ is a time-reversed version of a source with $\alpha = 0.6$, $S_{op} = 0.5$ and $S_{cp} = 0.6$ for any OQ value. While it is not yet clear what perceptual difference, if any, can be noticed between a source and its time-reversed variant, a simple analysis-by-synthesis test, in the time-domain, was employed at the end of the main algorithm to decide which version of the source should be selected.

It is important to note that while the proposed model is a flow model, the derivatives of the source signals are used in the construction of the codebooks to account for the radiation effects of the lips.

The constrained optimization on V_n to determine the vocal tract parameters require lower and upper bounds on the formant frequencies and their respective bandwidths. In [1], the bounds for formant frequencies were determined separately for male and female subjects by using the mean values as determined by the Snack Sound Toolkit [15]. However, the results were not consistent for two of the four subjects in that experiment, which estimated the parameter OQ. A possible explanation is the difficulty of estimating formant frequencies for the wide range of F_0 values, especially for phonations with high pitch. This is especially true for the first formant (F_1) frequencies which were found to have a large variance due to the well-known deficiencies of LPC-based formant trackers for high F_0 phonations. Since voice source characteristics are usually manifested in the lower frequencies of the speech spectrum, an inaccurate measurement of F_1 can lead to erroneous estimates of the source signal. In this study, two methods of determining formant frequency bounds are tested. The first method uses the Snack-estimated formant frequencies averaged across a subject's phonations and the second method uses formant frequencies which were manually extracted from the spectrum of a subject's normal phonation with normal F_0. Although the subjects were asked to produce the vowel /i/ for each recording, the end result was always different due to the positioning of the laryngoscope. For 3 male and 2 female subjects, the produced vowels are closer to an /ε/ vowel, while for the other female subject, the resulting vowels are closer to the /æ/ vowel. Using the Snack-estimated and manually-extracted formant frequency values, the lower and upper bounds for the constrained optimization were set to ± 150 Hz from the Snack/manual F_1 values, ± 250 Hz from the Snack/manual F_2 values, and ± 400 Hz for Snack/manual F_3 values. Table 1 shows the optimization constraints for the formant frequencies in terms of each subject for both methods; e.g. the Snack-estimated value for F_1 for subject FM1 is 351 Hz, therefore the lower optimization constraint is set to 201 Hz and the upper constraint set to 501 Hz. Bandwidth constraints are not shown here, but were based on the formant-bandwidth mapping formula in [16].

Table 1. *Optimization constraints for formant frequencies for each subject.*

	Snack-based lower/upper bounds (Hz)		
Subject	F_1	F_2	F_3
FM1	201/501	1466/1766	2116/2916
FM2	196/496	1331/1831	2437/3237
FM3	464/764	1454/1954	2550/3350
M1	287/587	1433/1933	2300/3100
M2	229/529	1310/1810	1999/2799
M3	176/476	1422/1922	2423/3223

	Manual-based lower/upper bounds (Hz)		
Subject	F_1	F_2	F_3
FM1	450/750	1430/1930	2115/2915
FM2	440/740	1650/2150	2350/3150
FM3	680/980	1620/2120	2550/3350
M1	380/680	1550/2050	2300/3100
M2	410/710	1350/1850	2000/2800
M3	380/680	1350/1850	1900/2700

4.2. Results and Discussion

Table 2 shows the results of the source estimation, for each phonation and F_0 type, in terms of the MSE averaged across the female and male subjects for the formant constraints as determined by Snack and manually. The MSE is calculated between the averaged source waveforms measured from the high-speed imaging and those estimated from the acoustic signals. In both cases, the averaged source waveforms are time and amplitude normalized so that each waveform is 100 samples in duration, and has a minimum value of 0 and a maximum value of 1. It can be seen that, the average MSE for most cases is lower for those sources which were estimated using manually-determined formant frequency constraints. Not surprisingly, the cases with high F_0 show the largest MSE difference between the Snack-based and manual-based methods, highlighting the inaccuracies of LPC-based formant estimators for high-pitched voices. To compare with the results in [1], which used the LF-model, outliers exist for a few cases. The Snack-based formant constraints method has significantly more errors which suggests that while the algorithm does not require precise formant estimates, the lower and upper bounds must contain a reasonably accurate formant position in order for the correct source to be selected from the codebook.

Table 2. *Results of the source signal estimation in terms of the MSE averaged across female and male subjects for a particular phonation and F_0 type. Both methods of determining the formant constraints are shown: Snack-based/manual-based.*

	Average MSE for female subjects (Snack/Manual)		
	low F_0	normal F_0	high F_0
pressed	.0783/.0113	.0173/.0945	.1276/.0277
normal	.0228/.0172	.0762/.0292	.1676/.1024
breathy	.0118/.0202	.0944/.0130	.1491/.0187

	Average MSE for male subjects (Snack/Manual)		
	low F_0	normal F_0	high F_0
pressed	.0216/.0216	.0057/.0065	.0935/.0382
normal	.0476/.0476	.0156/.0130	.0685/.0249
breathy	.0275/.0278	.0314/.0335	.1419/.0405

Table 3 shows the correlation coefficients resulting from the *OQ* estimation. With the exception of subject FM1, the correlation coefficients for "Manual" were similar or greater than those in our previous study [1]. While it maybe impractical to use manually-derived formant constraints in applications, a possible solution could be to use average formant values for known vowels (/æ/ for subject FM3 and /ε/ for the other subjects), as listed in [17]. The *OQ* estimation performance for these formant constraints is denoted by "Constant" in Table 3.

Table 3. *Cross correlation coefficients calculated between the estimated and measured OQ for each subject. "Previous" denotes the results from [1], "Manual" denotes the method using manually-based formant constraints and "Constant" denotes the formant constraints obtained from [17]. Note that only the first two female and male subjects were studied in [1].*

	Correlation coefficient (r) for each subject					
	FM1	FM2	FM3	M1	M2	M3
Previous	.971	.778	–	.925	.723	–
Manual	.913	.947	.865	.910	.919	.929
Constant	.892	.939	.925	.920	.919	.929

Visual inspections of the estimated source signal with the measured source signal confirm the results in Tables 2 and 3. For the manually-based formant constraints method, there are generally good matches between the estimated and measured signals, although

5. SUMMARY AND CONCLUSION

A new model for the voice source, based on observations from the high-speed imaging of the glottis, is proposed. This model is derived from the LF model by allowing greater flexibility in regards to the opening and closing phase durations. Although the proposed source model is able to closely match the glottal area waveforms as measured from imaging, all time-domain models lack an effective way of representing phonations with glottal gaps, i.e. DC-offsets. Incorporation of a noise component into the model may be a way to achieve this and forms the basis of future research.

The proposed source model is used in a source estimation algorithm based on the methods described in an earlier study [1]. While the results are quite promising, they show the importance of having reasonable formant frequency estimates, which can be difficult to obtain using LPC-based formant estimators. Sensitivity analysis of this algorithm will be further explored in future studies.

6. REFERENCES

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

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

[3] M. Howe and R. McGowan, "Sound generated by aerodynamic sources near a deformable body, with application to voiced speech," *J. Fluid Mech.*, vol. 592, pp. 367–392, 2007.

[4] A. Rosenberg, "Effects of the glottal pulse shape on the quality of natural vowels," *JASA*, vol. 49(2), pp. 583–590, 1971.

[5] G. Fant, J. Liljencrants, and Q. Lin, "A four-parameter model of glottal flow," in *STL-QPSR*, 1985, pp. 1–14.

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

[7] K. E. Cummings and M. A. Clements, "Glottal models for digital speech processing: a historical survey and new results," in *Digital signal processing*, 1995, vol. 5(1), pp. 21–42.

[8] G. Fant, *Acoustic theory of speech production*, Mouton, The Hague, Paris, 2nd edition, 1970.

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

[10] E. Moore II 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," *JASA*, vol. 110, pp. 479–488, 2001.

[12] P. Jinachitra and J. O. Smith III, "Joint estimation of glottal source and vocal tract for vocal synthesis using Kalman smoothing and EM algorithm," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2005, pp. 327–330.

[13] A. del Pozo and S. Young, "The linear transformation of LF glottal waveforms for voice conversion," in *Interspeech*, 2008, pp. 1457–1460.

[14] J. Pérez and A. Bonafonte, "Towards robust glottal source modeling," in *Interspeech*, 2009, pp. 68–71.

[15] Kåre Sjölander, "Snack Sound Toolkit," KTH Stockholm, Sweden, 2004, http://www.speech.kth.se/snack/ (last viewed Aug. 2009).

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

[17] G. E. Peterson and H. L. Barney, "Control methods used in a study of the vowels," *JASA*, vol. 24, pp. 175–184, 1952.