SOURCE MODEL ADEQUACY FOR PATHOLOGICAL VOICE SYNTHESIS

Melissa Epstein, Brian Gabelman, Norma Antoñanzas-Barroso, Bruce Gerratt, and Jody Kreiman
Voice Lab, Division of Head and Neck Surgery, UCLA School of Medicine, Los Angeles, USA 90095

ABSTRACT

Pathological voices are particularly difficult to inverse filter and fit with source models, in part, because of source-tract interactions inherent in these voice types. In order to obtain good synthetic copies of these voices, we need to know the practical importance of accurately inverse filtering and modeling individual voice sources. To this end, thirty different voices were synthesized, each with four different sources. Preliminary results suggest that the output of the inverse filter does not always contain enough information about the source to adequately reconstruct vocal quality. However, the LF model of the voice source pulse does provide enough degrees of freedom to model naturally-occurring quality variations across voices.

1. INTRODUCTION

Relatively few studies have examined synthesis of pathological voices. Historically, most synthesis attempts have been of normal male voices and synthesizer parameter development has reflected this. As speech synthesis has turned to modeling other phonation types, attention has also turned to better modeling of the voice source (e.g. Childers & Ahn [1]; Hillenbrand [2]).

Earlier studies, mainly by Childers and colleagues [1, 3, 4, 5, 6], have focused on obtaining average LF model source parameters for several phonation types (e.g. breathy, vocal fry, etc.). In these studies, natural voices were inverse filtered and the glottal source pulse was fit with an LF model. Parameters were averaged across several tokens of the same type of voice, and then speech samples and sentences were synthesized using increments of these averaged parameters. The resulting synthetic pathological voice samples varied in their perceptual acceptability.

At the same time, researchers have noted that it is difficult to inverse filter and model pathological voices because of source-tract interactions [1, 5, 7]. In particular, Childers & Ahn [1] noted that much of the variance in their average LF parameter values for pathological voices probably resulted from the inability of the LF model to accurately match the glottal pulses of the non-modal sources. They concluded, nonetheless, that the variances caused by errors in inverse filtering or LF modeling were unlikely to be a problem for speech synthesis.

Unlike previous studies, our work focuses on modeling a particular person's voice, and not just the general quality of breathiness, vocal fry, and so on. Ultimately, we want to be able to measure perceived voice quality – to parameterize the perception of pathological voices – without the use of dubiously valid perceptual terms. For example, synthesis can provide a tool through which listeners create a synthetic copy of a pathological voice that matches the sample under study, by manipulating synthesizer parameters. In this way, listeners' perceptions would be defined by the values of the synthesizer parameters they select and not by unstable internal models of the voice. Therefore, we need to know the practical importance of accurately inverse filtering and modeling individual voice sources. Considering that the LF model was developed primarily in the study of normal phonation, we also need to determine if the LF model is adequate for modeling the complexities of pathological vocal sources.

To this end, thirty different voices were inverse filtered using two different techniques (automatic and manual), and then synthesized with four different sources. We hoped to determine whether the perceptually important features of the pathological voice are preserved in inverse filtering and whether these sources can be adequately mimicked with an LF source model.

2. MODELING THE SOURCE

The LF model describes the glottal flow derivative with the combination of a sinusoid and an exponential [8]. Several points on the voice source pulse serve as parameters for minimizing errors between the model and the source under study. These points are usually the following major features of the glottal flow derivative (U´):

t_c = the length of the entire pulse
t_p = the length of the time that U´>0
t_e = the time of the maximum negative value of U´
E_e = the value of the maximum negative U´
t_a = the experimentally determined effective duration of the return phase

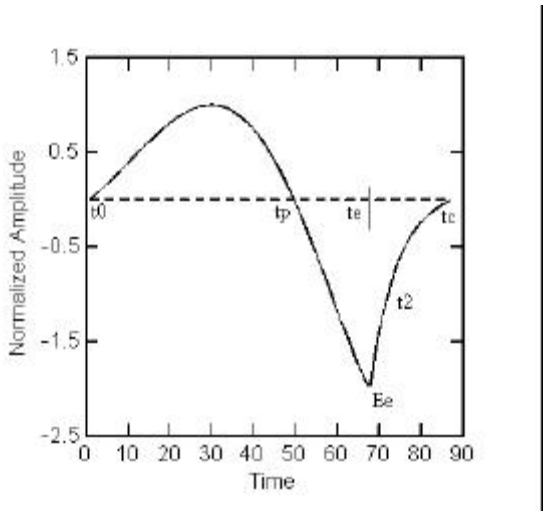

Figure 1. Revised LF model of differentiated glottal flow.

For our purposes, the traditional LF model has been modified in order to better fit pathological voices, particularly in the modeling of the return phase (Figure 1). T_a, the duration of

the return phase, has been replaced by t_2, the time increment to 50% decay in the return phase (i.e. where a line intersecting the halfway mark on the return phase intersects the time axis). A linear term is also added to the exponential segment, which is calculated to force the return phase back to zero. Finally, the equal area constraint was abandoned. Attempts to retain this constraint often resulted in source models that did not return to zero at the end of a pulse, resulting in discontinuities between pulses and high-frequency distortion in subsequent synthesis.

3. VOICE SYNTHESIS PROCEDURES

Thirty samples of the vowel /a/ were selected from randomly chosen speakers (15 male and 15 female) with voice disorders. Speakers ranged in age from 26 to 85. Diagnoses included vocal fold mass lesions, vocal fold paralysis, adductory spasmodic dysphonia, reflux laryngitis, glottal incompetence, and laryngeal web. Signals were transduced with a high-quality condenser microphone, sampled at 20 kHz and downsampled to 10 kHz.

A single cycle of phonation was chosen and concatenated to make a segment 10 cycles long. The voice source (flow derivative) and vocal tract transfer function were estimated based on inverse filtering and LPC/FFT analysis of this cycle. Flow derivative pulses were fitted with the modified LF model to produce several synthetic voice sources, as described below.

Once the glottal source and vocal tract transfer function were estimated, the noise spectrum was estimated by a cepstral-domain comb filter (similar to that described by de Krom [9]), which removed the harmonic part of the signal. The residual was then inverse filtered to remove the vocal tract parameters, leaving the inharmonic part of the source. This was smoothed and fitted with a 25 segment piece-wise linear approximation. (Note that jitter and shimmer were not modeled separately from the noise component.)

The synthesizer sample rate was fixed at 10kHz. A 100 tap FIR filter was synthesized for the noise spectrum, and a synthetic noise time series was created by passing white noise through this filter. Source pulses were added to the noise to create a complete glottal source time series. The ratio of noise to LF energy was adjusted so that the predicted noise-to-periodic energy ratio matched the value calculated from the original voice sample. Finally, the synthesized source was filtered through the vocal tract model to generate a preliminary version of the synthetic voice.

Within the synthesizer, the listener is free to adjust the following parameters: F0; formant values and bandwidths; noise level; tremor rate, excursion from mean F_0, and regularity; and the LF source parameters. These parameters were adjusted until the listener achieved the best synthetic voice possible, with the constraints on the particular source model described below. All parameters were adjusted separately for each source version. The time to complete analysis and re-synthesis of a single voice ranged from ten minutes to one hour.

4. EXPERIMENTAL DESIGN

Four different versions were created for each of the 30 target voices (Table 1). First, two different inverse filtering techniques were applied. In one, the inverse filtering algorithm ran automatically; in the second, an experienced operator (author JK) manipulated vowel formants and bandwidths to produce the smoothest possible output.

Both automatically and manually inverse filtered source pulses were then fitted with an LF model. Automatically inverse filtered sources were least-squares (LS) fitted with the LF model. Three different fitting approaches were applied to the manually filtered sources. In one version, the model was least-squares fitted to the flow derivative pulse. In the second, the LF parameters were visually fitted with the pulse, without regard to overall fit or to the perceptual adequacy of the result. A visual fit may do better than the least-squares minimization routine used to calculate the model if the least-squares calculations converged at a local minimum or did not run through enough iterations. Moreover, least-squares minimization seeks to match the overall shape of the glottal pulse, and not the location of the parameters. A visual fit allows for both.

	Inverse Filtering Condition	
	Automatic	Manual
LF Modeling Condition	LS fit	LS fit
		Major feature fit
		Perceptual fit

Table 1. Stimulus Modeling Conditions

In the third fitting method, the model parameters were manipulated to produce the best possible perceptual match to the target voice, without regard to the shape of the original volume velocity curve. In all cases, the values for all other synthesizer parameters were re-adjusted to compensate (as far as possible) for differences in the excitation provided by the various source models. All modeling was undertaken by the same experienced operator (author JK).

5. PERCEPTUAL EVALUATION

The following experiment was undertaken to examine the perceptual effects of the different inverse filtering and source modeling techniques applied above. Ten experienced listeners (speech pathologists, otolaryngologists, and linguists) participated in the experiment. They made eight comparisons of the different versions of each of the original thirty voices:

- the original target voice vs. each synthetic version;
- the perceptual fit vs. each of the other synthetic versions; and
- the automatically inverse filtered vs. the best possible inverse filtered result

for a total of 240 comparisons. Forty-eight pairs (20%; selected at random) were repeated to assess listener test-retest reliability. All stimuli were one second in duration. They were multiplied by 50 msec ramps to eliminate onset and offset click artifacts, and equalized for peak amplitude. Stimuli within a pair were separated by 500 msec.

Stimuli were presented by computer. Pairs of voices were randomized separately for each listener. Listeners were asked to listen carefully to each pair of voices and to rate their similarity by moving a cursor on a visual analog scale ranging from “exact

same" to "very different". Listeners controlled the rate of stimulus presentation and were able to replay voice pairs as desired before making their responses. Test time totaled approximately 40 minutes.

6. RESULTS

Detailed results will be presented at the conference. Preliminary data indicate that attempts to smooth the output of the inverse filter did not always improve the perceptual result relative to automatic inverse filtering. For many voices, operator intervention in the inverse filtering appears to have over-modeled the glottal pulse, smoothing away perceptually important details along with artifactual ripples. This resulted in a decrement in the perceptual match to the original voice. For other voices, the automatic inverse filter missed a pole in the higher region of the acoustic spectrum. This resulted in ripples in the output that made it difficult to fit the LF model to the flow derivative. Operator intervention in these cases improved the voice modeling process and the perceptual result.

Across conditions, the overall quality of the synthesis was judged to be excellent, with many voices nearly indistinguishable from the original samples. The overall good quality of the synthesis suggests that the LF model provides enough degrees of freedom to model naturally-occurring quality variations across the voices studied here.

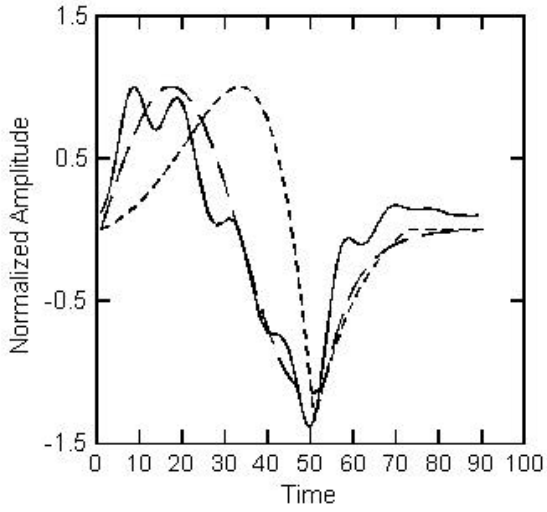

Figure 2. Three sources for a single voice. The solid line shows the (manually smoothed) output of the inverse filter; the long dashes show the least-squares fit of the LF model to the inverse filter output; the small dashes show the source that provided the best perceptual fit to the original voice sample.

Stimuli modeled with the perceptually best source were distinguishable from versions with other sources, and overall provided significantly better matches to the original voices. These sources sometimes differed substantially from the pulse derived by inverse filtering. For example, Figure 2 shows three source versions for a single voice. The solid line shows the output of the (manually smoothed) inverse filter; the long dashed

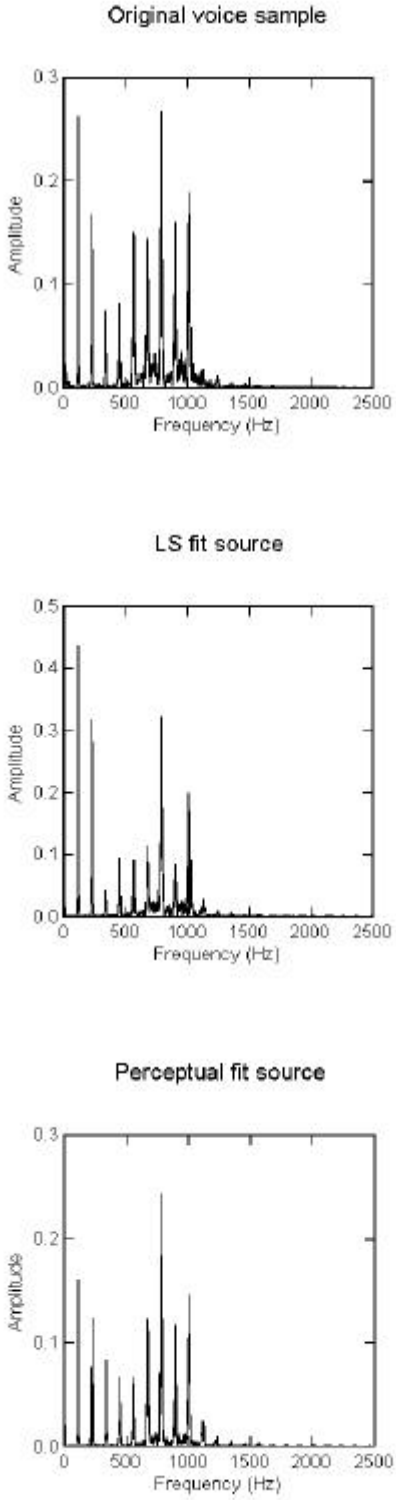

Figure 3. Linear FFT spectra for the original voice sample and for synthetic tokens created with the sources in Figure 2. Spectra have been truncated at 2.5 kHz to enhance detail.

line shows the least-squares fit of the LF model to that output; and the small dashed line shows the source that provided the best perceptual fit to the original voice sample. The original voice sample and synthesized vowels created with the two smoothed sources are attached as sound files [0415.wav] (the voices are, in order, the original, the least-squares fit, and the perceptual fit). Figure 3 shows linear FFT spectra for the original voice and these two synthetic versions. Despite many attempts at modifying spectral noise and formant bandwidths, it was impossible to achieve a good spectral match to the original voice with the least-squares fit source. A much better match was obtained when the source pulse was modified as shown in the figure. This case and others like it suggest that the output of the inverse filter does not always contain enough information about the source to adequately reconstruct vocal quality. Thus, further research is needed to examine the extent to which the output of the inverse filter can be used as a model of the vocal source.

7. CONCLUSIONS

A. Extensive operator intervention in the inverse filter does not appear to improve the resulting voice synthesis, and may even reduce the match of the copy to the original.

B. The output of the inverse filter – whether smoothed or not – does not necessarily provide a good template for a perceptually adequate source model.

C. Good perceptual matches to the original stimuli can be created with the LF source model, although the model that provided a good match did not necessarily resemble the flow derivative waveform particularly closely. This suggests that source models more complex than the LF model are not required to model the variations in voice quality that occur with vocal pathology.

ACKNOWLEDGEMENTS

This research was supported in part by NIH grant DC 00029-09 and DC 01797. We would like to thank Donald Dirks and Sumiko Takayanagi for their assistance.

REFERENCES

[1] Childers, D.G. and Ahn, C. 1995. Modeling the glottal volume velocity waveform for three voice types. *Journal of the Acoustical Society of America*, 97, 505-519.

[2] Hillenbrand, J. 1988. Perception of aperiodicities in synthetically generated voices. *Journal of the Acoustical Society of America*, 83, 2361-2371.

[3] Childers, D.G. 1995. Glottal source modeling for voice conversion. *Speech Communication*, 16, 127-138.

[4] Childers, D.G. and Lee, C.K. 1991. Vocal quality factors: Analysis, synthesis and perception. *Journal of the Acoustical Society of America*, 90, 2394-2410.

[5] Childers, D.G. and Wong, C. 1994. Measuring and modeling vocal source-tract interaction. *IEEE Transactions on Biomedical Engineering*, 41, 663-671.

[6] Lalwani, A.L. and Childers, D.G. 1991. Modeling vocal disorders via formant synthesis. *Proceedings of the IEEE*, 505-508.

[7] Strik, H. 1998. Acoustic parameterization of differentiated glottal flow: Comparing methods by means of synthetic flow pulses. *Journal of the Acoustical Society of America*, 103, 2659-2669.

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

[9] de Krom, G. 1993. A cepstrum-based technique for determining a harmonics-to-noise ratio in speech signals. *Journal of Speech and Hearing Research*, 36, 254-266.