Analysis of Nonmodal Phonation using Minimum Entropy Deconvolution*

Nicolas Malyska and Thomas F. Quatieri

MIT Lincoln Laboratory
MIT Lincoln Laboratory, Lexington, MA USA
{nmalyska, quatieri}@ll.mit.edu

Abstract

Nonmodal phonation occurs when glottal pulses exhibit non-uniform pulse-to-pulse characteristics such as irregular spacings, amplitudes, and/or shapes. The analysis of regions of such nonmodality has application to automatic speech, speaker, language, and dialect recognition. In this paper, we examine the usefulness of a technique called minimum-entropy deconvolution, or *MED* [1], for the analysis of pulse events in nonmodal speech. Our study presents evidence for both natural and synthetic speech that MED decomposes nonmodal phonation into a series of sharp pulses and a set of mixed-phase impulse responses. We show that the estimated impulse responses are quantitatively similar to those in our synthesis model. A hybrid method incorporating aspects of both MED and linear prediction is also introduced. We show preliminary evidence that the hybrid method has benefit over MED alone for composite impulse-response estimation by being more robust to short-time windowing effects as well as a speech aspiration noise component.

Index Terms: inverse filtering, nonmodal speech, glottal pulse, minimum entropy

1. Introduction

Observations of *nonmodal* phonation, with glottal pulses having non-uniform characteristics such as irregular spacings and amplitudes have been often reported in speech science literature. Instances of this class of phonation have been referred to using terms such as "creak," "vocal fry," "diplophonia," "irregularity," and "glottalization." In this paper, we study a method of analyzing individual glottal events, which are common underlying units of these phonation types. Figure 1 shows a nonmodal phonation manifesting at the end of a natural utterance.

We are motivated to study nonmodal phonation for several reasons. First, nonmodal phonation is a common occurrence in the speech of both normal and disordered speakers. One study investigating this type of phonation has reported that a set of normal speakers exhibited nonmodal behavior for between 13 and 44 percent of their word-initial vowels [2]. Another reason to study nonmodal phonation is that the timing, amplitude, and other characteristics of individual glottal pulses during nonmodal regions may be dependent upon linguistic cues, speaker identity, language, and dialect. Pulse characteristics may also be dependent on vocal-fold pathologies.

A common model for the generation of voiced speech is a volume-velocity source waveform, filtered by both an all-pole minimum-phase vocal tract filter and a radiation characteristic at the mouth to produce an acoustic pressure signal. As an additional step, each source pulse can be modeled as a pure impulse source convolved with a mixed-phase *source response* [3]. In this paper, we call the mixed-phase combination of the source response, vocal tract, and radiation characteristic the *composite impulse response*. Our primary goals are to (1) estimate the composite impulse response of a given speech region and (2) generate the source impulse sequence using an inverse filter derived from the composite impulse response. One motivation for obtaining this representation is the use of timings and amplitudes of the source impulse sequence in automatic speech, speaker, language, and dialect recognition systems.

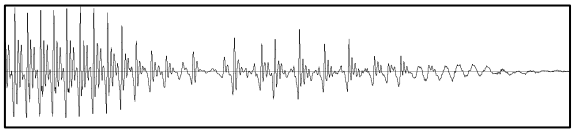

Figure 1. A section of natural nonmodal phonation from the end of the word "umbrella" in a normal speaker.

In particular, we examine the application of an algorithm called minimum-entropy deconvolution (MED) [1] to the problem of determining the composite impulse response and associated source impulse sequence in sections of nonmodal phonation. The use of MED on near-modal speech to derive a pulse-like residual has been studied previously with promising results [4]. We build upon the previous work by showing MED's application to deriving a pulse-like signal from highly-nonmodal synthetic and natural phonation and to estimating composite impulse responses. In addition, we propose a hybrid method that combines MED with conventional linear prediction. Evidence is presented that this hybrid method has benefit over MED alone for composite impulse-response estimation by being more robust to a speech aspiration noise component as well as to the effects of short-time windowing.

2. Minimum-Entropy Deconvolution

In this paper, we assume that nonmodal phonation is the result of convolving a source impulse sequence with a composite impulse response. A reasonable approach to the decomposition problem, then, is to design an inverse-filtering method that yields an impulse-like residual. An alternative to linear prediction, a traditional method for deriving such a pulse-like residual, is a technique from the geophysical literature called minimum-entropy deconvolution (MED). This method contrasts linear prediction primarily in two ways. First, while linear prediction maximizes the entropy in the residual, MED attempts to minimize disorder. In effect, it creates a filter that generates the most "pulse-like" residual for a given input. The second way in which MED and linear

*This work is sponsored by the United States Air Force Research Laboratory under Air Force Contract FA8721-05-C-0002. Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the United States Government.

prediction differ is that the MED filter is able to generate pure impulses from a mixed-phase system as from a composite speech impulse response. Inverse filtering with linear prediction, on the other hand, can create an impulse sequence for only a minimum-phase system.

MED achieves minimum entropy in its output by solving for a set of filter coefficients that maximize a criteria of pulse-likeness called the *varimax norm* [1]:

$$V = \sum_j x^4[j] \Bigg/ \left(\sum_j x^2[j] \right)^2 \quad (1)$$

where $x[j]$ is the input signal. The varimax norm is a kind of normalized 4th-order moment, *kurtosis*, and is higher for signals with a small number of sharp pulses and closer to zero for signals with less structure. The problem of maximizing the varimax norm of the residual is nonlinear and requires an iterative approach as detailed in [1] and [4].

3. Application of MED to Nonmodal Speech

MED may be applied to continuous speech using analysis with overlapping windows. In our implementation, we process the speech using 20-ms Hamming windows with 50-percent overlap. Creation of the residual from MED filters is performed using frame-wise analysis and an overlap-and-save synthesis approach [5]. Our full implementation of such a system with MED is beyond the scope of this paper, and does not yet exist in the literature. Analysis using a single large analysis window is described in [4].

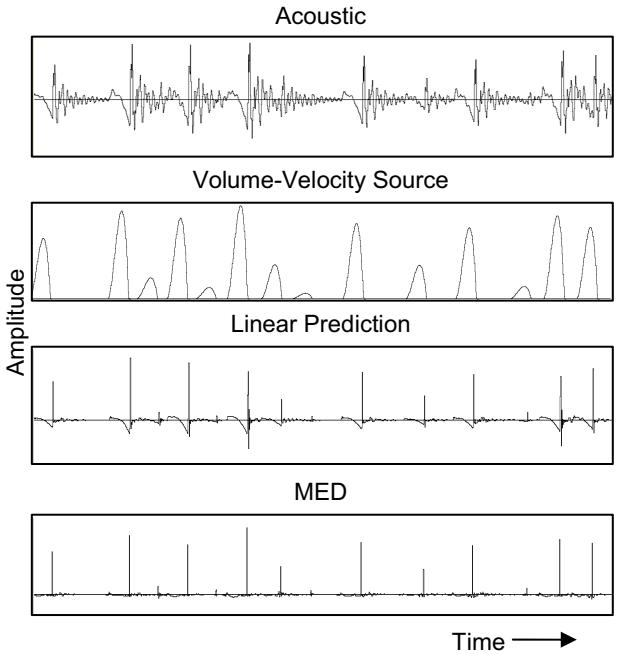

Figure 2. Application of MED to synthetic nonmodal phonation. The bottom two panels show the application of the algorithm to the acoustic wave in the top panel. Linear prediction results are shown for comparison.

Figure 2 shows the results of applying MED to synthetic nonmodal phonation. The top two panes depict the synthetic volume-velocity source and resulting acoustic signal. This particular case contains source pulses with both nonuniform interpulse timings and amplitudes. The two remaining panes show the results of using a 15th-order linear prediction and a 25th-order MED to obtain a residual. These values were used because increasing beyond order 15 and 25 for linear prediction and MED, respectively, gave negligible performance gain. The output of linear prediction is less pulse-like than the output of MED, with activity to the left of the primary output impulses. In contrast, MED yields a qualitatively more impulse-like residual; there is less activity to the sides of the main pulses. In both examples, the most impulse-like activity corresponds to near the closing point of the synthetic volume-velocity source. Figure 3 repeats this comparison for a natural nonmodal utterance. Although there exists no standard "ground truth" for pulse locations in natural nonmodal phonation, we see that MED is again qualitatively more pulse-like.

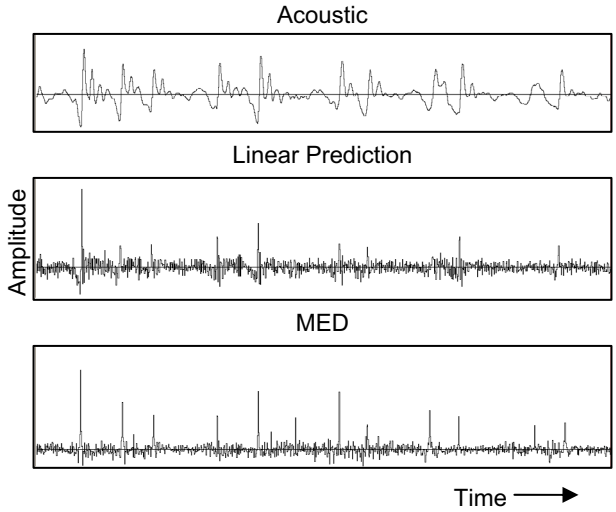

Figure 3. Output of MED compared with the output of linear prediction for a natural speech signal. The MED residual is more pulse-like than the linear-prediction residual.

Using the varimax norm as an objective measure of pulse-likeness, we find that MED is the most pulse-like compared with linear prediction and the windowed input waveform for both the natural and synthetic cases. Across 11 10-ms frames of the natural nonmodal phonation in Figure 3, MED has an average varimax norm of 0.24 versus 0.06 for linear prediction and 0.04 for the input. For 126 frames of the synthetic case in Figure 2, MED has an average varimax norm of 0.66 versus 0.34 for linear prediction and 0.04 for the input.

We can also calculate how well the estimated composite impulse responses from MED and linear prediction fit the true composite impulse response using the mean-squared error. Since the true impulse-response is not known for the natural speech, we can only perform this test for synthetic utterances. For 126 frames of synthetic nonmodal phonation of Figure 2, MED yields a closer fit to the known composite impulse response. The mean-squared error is on average 0.18 versus 0.44 for linear prediction.

4. Hybrid Linear-Prediction/MED Approach

We have shown that MED produces a higher varimax norm and a closer fit to a known composite impulse-response than linear prediction, but this result is not surprising. The

composite impulse response we aim to recover is represented well as the output of a mixed-phase system [3, 6]. Linear prediction effectively "flips" maximum-phase pole estimates to their minimum-phase reciprocal locations. In this section, we present a modification of standard linear prediction capable of obtaining an improved mean-squared error fit to mixed-phase impulse responses over linear prediction.

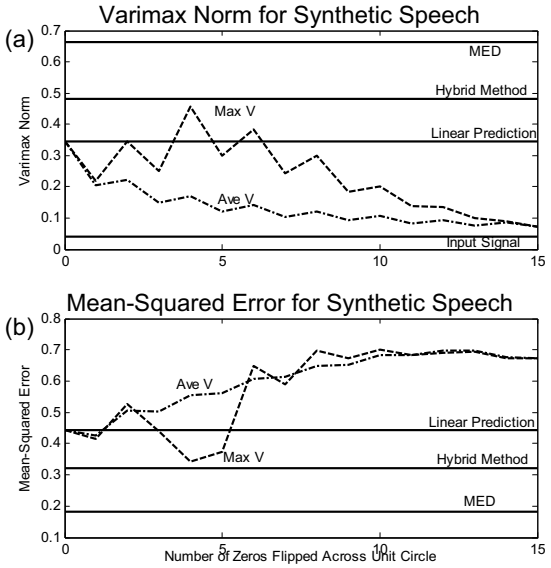

Figure 4. Average (a) varimax norm and (b) mean-squared error measurements across 126 input frames of synthetic nonmodal phonation. Solid horizontal lines indicate linear prediction, hybrid method, and MED measurements. Dashed and dash-dot lines show results for the maximum and average varimax norm respectively for each number of flips of linear-prediction-derived zeros across the unit circle.

The least-mean-square residual criteria upon which linear prediction is based cannot be used to determine which location for each pole—inside or outside the unit circle—is "better." Instead we need an additional criterion. We propose to use the varimax norm criteria from MED to choose among the possible configurations. In this way, a filter is created that yields a linear-prediction-based residual with the maximum pulse-likeness possible. We will show that this combination allows us to estimate realistic composite impulse response shapes. For a specified order, the solution we use is to (1) try every possible configuration of zeros inside and outside the unit circle, (2) find the residual resulting from each configuration, and (3) keep the configuration yielding the maximum varimax norm. The complexity of this technique grows exponentially with increasing linear-prediction order, but is manageable for typical orders. We show here results for an order-15 case.

Figure 4(a) compares the varimax norm of the input signal to the residuals using linear prediction, MED, and the hybrid method for 126 frames of the synthetic nonmodal phonation example of Figure 2. The results show that the MED residual is the most pulse-like on average out of all of the methods followed by the hybrid method, linear prediction, and the input waveform. In this figure, we also illustrate how different "flips" of the linear prediction zeros affect the varimax norm. The average (dash-dot line) and maximum (dashed line) values of the varimax norm are plotted for each number of zeros flipped across the unit circle to their reciprocals. We can see that, in this example, the maximum varimax norm comes from flipping four zeros outside of the unit circle. Figure 4(b) compares the mean-squared errors of the best fits between the composite impulse response derived using each method and the known impulse response. MED again yields the lowest mean-squared error fit followed by the hybrid method and linear prediction. We can see from the dashed curve that out of all the linear-prediction zero configurations, the number of zeros outside the unit circle yielding the largest varimax norm also yields the smallest mean-squared error. This finding supports the idea that a pulse –like residual is an important indicator of a good estimate of the composite impulse response.

5. Further Evaluation of Composite Impulse Response Estimates

Figure 5 compares the composite impulse responses derived from each of the methods for examples of both natural and synthetic speech. In both cases, the hybrid approach is seen to qualitatively provide advantages over the conventional linear-prediction output, for example allowing representation of the negative peak in the impulse response. The MED impulse-

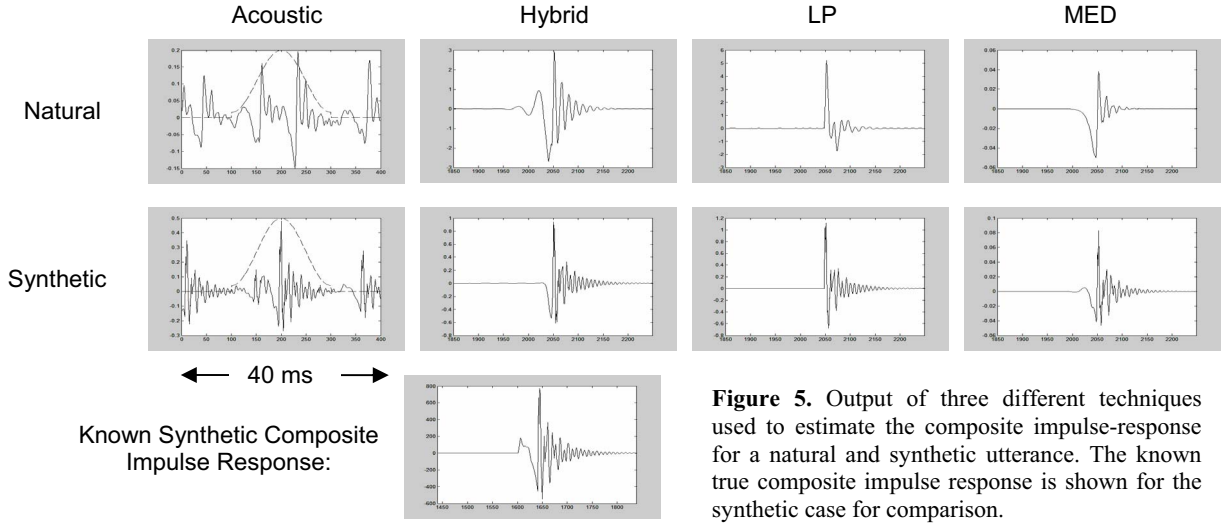

Figure 5. Output of three different techniques used to estimate the composite impulse-response for a natural and synthetic utterance. The known true composite impulse response is shown for the synthetic case for comparison.

response estimate appears to have an improved fit over linear prediction for the real and synthetic cases. Despite producing estimates like the one shown for most frames, however, MED has a tendency to occasionally yield unrealistic composite impulse-response shapes, even for our synthetic example. An example of this phenomenon is shown for synthetic speech in Figure 6. While the hybrid method yields a reasonable impulse response, which decays in about 20 ms, MED estimates a composite impulse response that rings significantly for over 100 ms. As can be seen in the acoustic waveform, the pulses in this example occur close to the edges of the analysis window. This distorts the waveform that MED and linear prediction process. Based on this and other examples, it appears that the hybrid method is not prone to the unrealistic "ringing" behavior that can occur with MED.

The ringing behavior seen in the previous example occurs in less than 5 percent of the synthetic speech frames analyzed and is likely due to the location of the excitation near the edges of the window. We have further evidence, however, that MED is overall not as robust as the hybrid method when confronted with real speech. Subjectively, the composite impulse response estimates contain ringing and other problems more often for real speech. Toward understanding this sensitivity, we simulate one perturbation typical of real speech—increased aspiration noise—in our synthetic utterance, and visualize how the average of our three methods respond to this change.

Figure 7 shows results from synthetic nonmodal phonation with added aspiration noise varied from 0 dB to 45 dB to 60 dB. In estimating the composite impulse response, MED appears to be significantly more sensitive to the increased aspiration noise, yielding a change in average mean-squared error of more than 0.6. In contrast, the error in the hybrid method changes by only about 0.35. In the varimax-norm plot, it can be observed that there is a decrease for all three methods as the amount of aspiration noise is increased.

At this point in the research, we do not completely understand why MED is sensitive to deviations from the assumed model and to the position of the signal in the window. These problems, however, illustrate that the hybrid method can produce an improved composite impulse response over MED under certain conditions.

6. Conclusions

In this paper, we have shown the result of applying MED to the problems of decomposing nonmodal phonation into a pulse-like residual and a realistic composite impulse response. In particular, we have shown evidence that both MED and a hybrid method combining linear prediction and MED can be used to obtain pulse-like residuals and accurate composite impulse responses. Although MED on average produces both the most pulse-like residual and the best impulse-response fit for synthetic data, we have shown evidence that it is not as robust as the hybrid method. In particular, it occasionally produces composite impulse responses that "ring" excessively and is also sensitive to deviations of the underlying pulse-excitation model on which it is based. In future work, we plan to combine MED and the hybrid method to exploit their individual strengths and obtain more robust estimates of the composite impulse response.

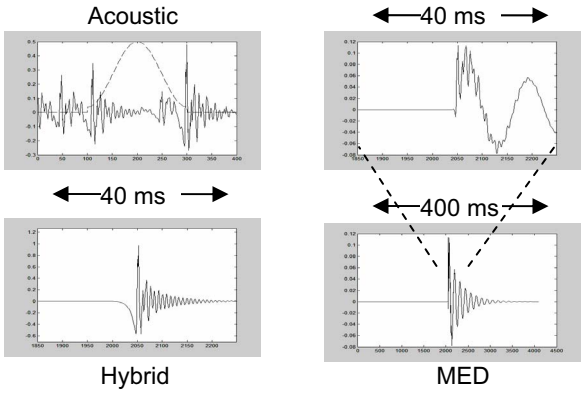

Figure 6. Application of MED and the hybrid method to a frame of synthetic nonmodal phonation that yields an unrealistic estimated glottal impulse response for MED. The MED impulse response "rings" for close to 200 ms as can be seen in the bottom-right panel.

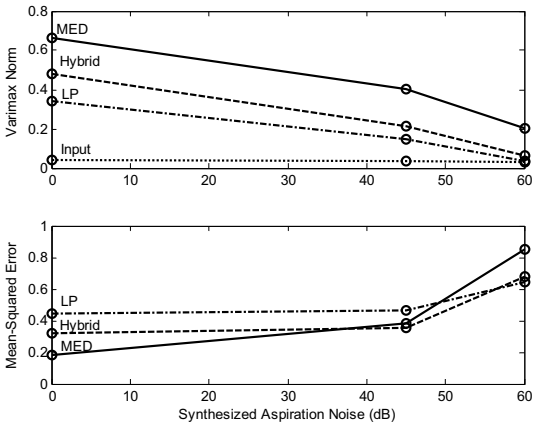

Figure 7. Comparison of mean varimax norm and mean-squared error in the composite impulse response for three different levels of synthetic aspiration noise across 126 input frames.

7. Acknowledgements

The authors would like to thank Janet Slifka and Stefanie Shattuck-Hufnagel for access to their data containing nonmodal regions.

8. References

[1] R. A. Wiggins, "Minimum entropy deconvolution," *Geoexploration*, vol. 16, pp. 21-35, 1978.

[2] L. Dilley, S. Shattuck-Hufnagel, and M. Ostendorf, "Glottalization of word-initial vowels as a function of prosodic structure," *Journal of Phonetics*, vol. 24, pp. 423-444, 1996.

[3] T. F. Quatieri, *Discrete-Time Speech Signal Processing: Principles and Practice*. Upper Saddle River, NJ: Prentice Hall, 2002.

[4] G. Gonzalez, R. E. Badra, R. Medina, and J. Regidor, "Period estimation using minimum entropy deconvolution (MED)," *Signal Processing*, vol. 41, pp. 91-100, 1995.

[5] A. V. Oppenheim, R. W. Schafer, and J. R. Buck, *Discrete-Time Signal Processing*, 2nd ed. Upper Saddle River, N.J.: Prentice Hall, 1999.

[6] M. R. Matausek and V. S. Batalov, "A new approach to the determination of the glottal waveform," *IEEE Transactions on Acoustics Speech and Signal Processing*, vol. 28, pp. 616-622, 1980.