SYNTHESIS, ANALYSIS, AND PITCH MODIFICATION OF THE BREATHY VOWEL*

Daryush Mehta and Thomas F. Quatieri

MIT Lincoln Laboratory
244 Wood Street, Lexington, MA 02420
[dmehta, quatieri]@ll.mit.edu

ABSTRACT

Breathiness is an aspect of voice quality that is difficult to analyze and synthesize, especially since its periodic and noise components are typically overlapping in frequency. The decomposition and manipulation of these two components is of importance in a variety of speech application areas such as text-to-speech synthesis, speech encoding, and clinical assessment of disordered voices. This paper first investigates the perceptual relevance of a speech production model that assumes the speech noise component is modulated by the glottal airflow waveform. After verifying the importance of noise modulation in breathy vowels, we use the modulation model to address the particular problem of pitch modification of this signal class. Using a decomposition method referred to as pitch-scaled harmonic filtering to extract the additive noise component, we introduce a pitch modification algorithm that explicitly modifies the modulation characteristic of this noise component. The approach applies envelope shaping to the noise source that is derived from the inverse-filtered noise component. Modification examples using synthetic and real breathy vowels indicate promising performance with spectrally-overlapping periodic and noise components.

1. INTRODUCTION

This paper investigates the synthesis and analysis of the breathy vowel with application to pitch modification. The approach builds on sinusoidal modeling of speech [1] and various extensions that decompose the speech signal into periodic and noise components [2-4]. Certain extensions are based on an additive noise model that is *modulated* at the pitch and synchronized with the voiced component [3, 4]. A challenge for analysis based on this model is the accurate separation of a noise component with not only the correct temporal modulation and synchronization but also the correct spectral characteristics that typically overlap the periodic component.

Although standard decomposition methods used in speech modification assume the noise and periodic signal components spectrally overlap, they ultimately simplify the analysis with a noise component that lies in a spectrally disjoint high-frequency region [3, 4]. Recently, however, a number of decomposition techniques have been introduced that show improved accuracy in getting at modulated noise that is spectrally mixed with its periodic companion [5, 6]. In this paper, we select one of these techniques, pitch-scaled harmonic filtering [5], as a basis for periodic/noise component decomposition. We then use this decomposition as a front-end to a pitch modification algorithm that alters the noise component in a manner consistent with the modification of the periodic component. This consistency is achieved by altering the envelope modulation of the noise source derived from inverse filtering the speech noise component. Modifying the envelope of the noise source, which is a function of the glottal flow waveform, more closely matches actual production of modified pitch than altering the envelope of the dispersed noise at the output of the vocal tract, as done in previous approaches [3, 4]. High-quality pitch modification of speech is desirable in numerous applications, including text-to-speech synthesis and clinical tools for assessing voice disorders. Controlling pitch and noise parameters may allow clinicians to modify disordered breathy voices and estimate improvements in speech acoustics after patients undergo voice therapy or surgery.

This paper is organized as follows. Section 2 briefly reviews the physiology of the breathy voice that leads to the modulation model of the speech noise component. Section 3, motivated by the work of Hermes [7], explores the naturalness of vowel synthesis using various types of noise source envelopes to understand their perceptual importance. Specifically, we look at constant, sinusoidal, and glottal-flow amplitude modulation applied to white Gaussian noise. In Section 4, we next describe the pitch-scaled harmonic filtering (PSHF) method [5] to separate the periodic and noise components of a breathy vowel. In Section 5, we then use the output of the PSHF method as input to our pitch modification algorithm. We apply standard sinewave-based modification on the periodic component [1], while for the noise component we derive a noise source envelope with a new modulation rate synchronized with the modified periodic component. Finally, in Section 6, we summarize features of the pitch modifier and discuss next steps.

2. PRODUCTION MECHANISM AND MODEL

Typically the term *breathiness* refers to turbulence generated at the glottal level, acting as a noise source to the vocal tract. High-velocity air pushes through the glottal constriction and

* This work was sponsored by the Department of Defense under Air Force Contract F19628-00-C-0002. Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the United States Government.

results in the generation of a jet stream with eddies of air anterior to the constriction that introduce noise into the production of speech [8]. In some cases, the turbulence generates pressure sources that may be distributed over several regions near the glottis [9]. More generally, however, turbulence can be created at other locations along the vocal tract away from the glottis, as with voiced and unvoiced fricatives. Although, in these cases, the voice is not necessarily breathy, a noise component is introduced at the vocal tract output.

When the vocal folds are vibrating during the generation of turbulence, the resulting noise source is said to be *modulated by the glottal flow volume velocity* with a larger pressure source resulting from higher-velocity turbulences [9]. This modulated noise source is thought to occur in breathy vowels, voiced fricatives, and many forms of vocal dysphonia. In addition, in the posterior part of the vocal folds near the arytenoid cartilages, a posterior glottal opening or chink is often present, allowing a constant DC flow of air during phonation [10]. The resulting air flow turbulence is generated at or near the glottis, aiding in creating the breathy percept [11].

In this paper, our focus is on the breathy vowel, although our approach is more general. Figure 1 shows the model used for synthesizing a breathy vowel, based on the above observations. The speech waveform consists of a linear combination of periodic and noise components. The periodic component is the output of a linear vocal tract filter with a periodic glottal flow velocity source, while the noise component is the output of the vocal tract filter with its input being a noise source modulated by the glottal flow volume velocity waveform. Both components are shaped by a radiation characteristic.

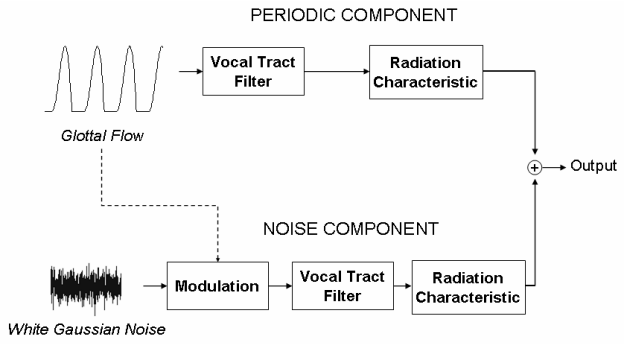

Figure 1: *Model of the breathy vowel*

Figure 1 presents this speech production model, simulated in Matlab with the derivative of the glottal flow (radiation taken into account) implemented using a Liljencrants-Fant (LF) model [12] and the vocal tract transfer function derived from a 3-pole model implemented as a cascade of digital resonators.

3. PERCEPTION OF MODULATION

Using our simulation of Figure 1, and motivated by the earlier work of Hermes [7], we have explored the noise modulation implemented in the modulation block of Figure 1. First, preliminary informal listening supports Hermes' finding that there is a perceptual difference between synthetic vowels with noise modulated at the pitch and those with unmodulated noise. Moreover, also as found by Hermes [7], the noise integrates with the periodic component best when the modulated noise is synchronized with the periodic component. These results indicate that synchronized modulation is important for naturalness and perceptual fusion but do not reveal how best to select the modulation function.

To explore this question, we investigated four different modulation patterns illustrated in Figure 2: rectangular (no modulation), sinusoidal, and periodic glottal volume velocity with and without a DC component. Informal listening indicates that the glottal airflow waveform provides for the most natural synthesis, with the DC component addition slightly preferred, consistent with our assumed modulation model described in Section 2. As the vocal folds oscillate, they open and close to effectively gate any generated noise, while the possible presence of a posterior glottal chink serves as a source of constant (DC) flow of air.

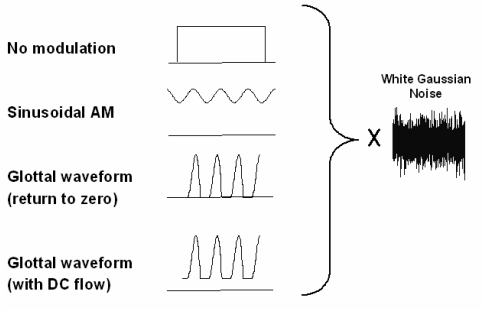

Figure 2: *Different noise modulation envelopes*

4. ANALYSIS OF A BREATHY VOWEL

A recent decomposition technique, pitch-scaled harmonic filtering (PSHF) [5], was implemented in Matlab to separate the periodic and noise components of a breathy vowel. The PSHF method is desirable since there is evidence that it can preserve the temporal modulation characteristics of the noise component. The PSHF approach uses an analysis window duration equal to four pitch periods and relies on the property that harmonics of the fundamental frequency fall at specific frequency bins of the short-time Fourier transform (see [5] for details). Spectral subtraction is subsequently performed to obtain the noise component spectrum. Pitch periods are estimated using the speech signal processing tool Praat [13].

This decomposition technique approximately isolates the noise component in a breathy utterance. Some small leakage of harmonicity can be present in the extracted noise component. In spite of this limitation, we decided to proceed with using the algorithm as a front-end decomposition for the proposed pitch modification scheme because it provides the general signal characteristics of interest.

5. MODIFYING A BREATHY VOWEL

With knowledge of the perceptual importance of modulated noise in a speech signal and a means to approximately decompose spectrally overlapping periodic and noise signal

components, we turn to describing an algorithm for modifying the pitch of a breathy vowel, accounting for a modulated noise source. We first briefly review a sinewave-based pitch modifier that forms the baseline for our modification strategy.

5.1. Sinewave-based pitch modification

Sinusoidal analysis/synthesis is based on the premise that speech can be modeled as a sum of sinewaves with time-varying amplitudes, frequencies, and phases [1]. Sinusoidal modification uses an analysis/synthesis strategy based on this sinewave model. Speech transformations stretch and compress sinewave frequency trajectories in time and frequency for time-scale and pitch-scale modification, respectively.

Because the basic modification scheme does not decompose the noise component from the periodic component, there is little control over the resulting noise signal by a system tuned to modifying the periodic component. Consequently, a number of extensions have been developed to account for an additive modulated noise component [3, 4]. As alluded to earlier, although these decomposition methods assume the noise and periodic signal components spectrally overlap, they ultimately simplify the analysis with a noise component that lies in a spectrally disjoint high-frequency region. In addition, although these methods have recognized the importance of preserving noise modulations in signal modification, the envelope of the noise at the vocal tract output is modified and not the envelope of the noise at the source, which would be more consistent with the modulation model of Figure 1.

5.2. Pitch modification of a vowel with modulated noise

In our pitch modifier, the PSHF technique eliminates the need of setting a frequency boundary between periodicity and noise, while inverse filtering allows for operating on the noise source itself. In other words, we perform inverse filtering prior to envelope estimation so we may modify the modulations as they appear prior to vocal tract filtering. Using the original noise source aims to preserve the temporal character of the speech signal.

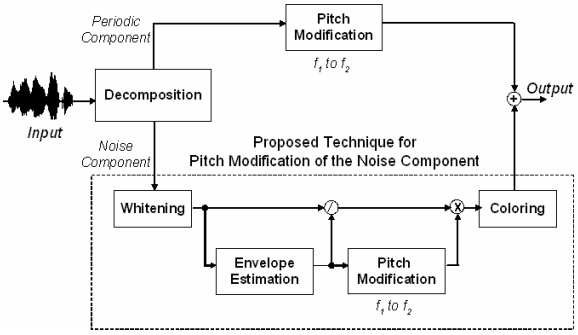

Figure 3: *Pitch modification algorithm*

As illustrated in Figure 3, the input speech noise component is first sent through a periodic/noise decomposition block, consisting of the PSHF algorithm. The periodic and noise components are then modified separately and summed. The periodic component is modified with the sinewave modification system [1], whereas the noise component is separately modified with the following steps:

Whitening: This operation utilizes a short-time whitening filter. A 20-ms Hanning analysis window is applied with a half-window overlap. Whitening is accomplished in each windowed short-time segment by the following algorithm:

- Linear predictive estimation of the all-pole model representing the vocal tract filter formant frequencies
- Inverse filtering of the short-time segment by an FIR filter whose tap weights are the corresponding coefficients in the estimated all-pole model
- Overlap-and-add synthesis of the spource signal

Figure 4 compares the waveform of a breathy synthetic vowel /a/ and the output of the whitening block. This test vowel has three formants (820, 1220, and 2810 Hz), the DC glottal flow parameter is zero, and the open quotient is 0.6. The sampling rate is 8 kHz and the vowel's fundamental frequency is 125 Hz.

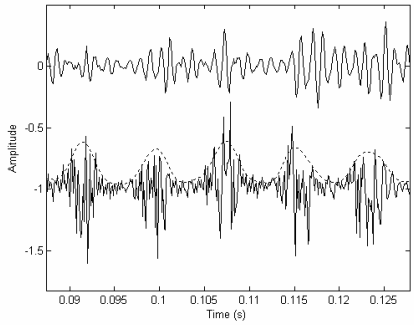

Figure 4: *Process of whitening the noise component (top) Estimated noise component (bottom) The solid line is the inverse filtered noise signal, i.e., the estimated noise source, and the dotted line corresponds to its estimated envelope*

Envelope estimation: The result of this block uncovers the noise modulations generated by the glottal airflow undulations. We have evaluated several envelope detection algorithms, including the Hilbert transform and the demodulation process of half-wave rectification followed by low-pass filtering. We have chosen to perform envelope estimation using a combination of the Hilbert transform and low-pass filtering. The magnitude of the Hilbert transform is calculated and filtered by a 50-tap FIR filter with a cutoff frequency of 350 Hz (since pitches observed are less than 300 Hz). For the vowel /a/, the estimated envelope of a section of estimated modulated noise source is shown in Figure 4 (bottom waveform).

Pitch Modification: Recall the model of breathy vowel production in Figure 1. The pitch modification algorithm uncovers the periodicity in the production of the noise component and re-modulates the glottal noise source with a new pitch-scaled envelope. To perform this envelope modification, in our preliminary work, resampling is performed on the envelope of the whitened modulated noise signal. A comparison between the estimated noise source component before and after pitch modification is presented in Figure 5 with the above breathy vowel /a/ as the example.

Coloring: This block spectrally colors the newly-modulated noise component with the parameters of the vocal tract filter estimated in the whitening block.

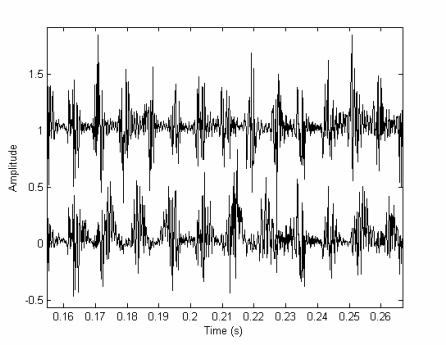

Figure 5: *Pitch modification of noise source from 125 Hz to 100 Hz. (top) Whitened signal and (bottom) re-modulated noise source before spectral coloring.*

5.3. Perceptual experiments

In lowering pitch of synthetic and real vowels from 125 to 100 Hz, with informal listening, we evaluated our pitch modifier and compared with the sinewave modification baseline.

With synthetic vowels /a/, /i/, /e/, /o/, and /u/ as input to our proposed algorithm, listeners typically describe the output as having voiced and noise components perceptually integrated. The modified noise was not heard as a distinct entity. With sinewave-based modification, on the other hand, the noise component is perceived as somewhat tonal and more perceptually separate from the periodic component. We also pitch-modified samples of the breathy vowel /a/ selected from a database of pathological voices [14]. When the breathy vowel of a pathological speaker is pitch-modified using the proposed algorithm, the output signal is typically perceived to take on a naturally breathy quality. Comparison with signals pitch-modified by sinewave modification is essentially consistent with that from synthetic signal modification, but the decomposition-based modification is somewhat less effective because real vowel characteristics tend to deviate from the ideal harmonic structure of synthetic vowels.

6. CONCLUSIONS

In this paper, we have proposed a pitch modification algorithm that specifically seeks to modify the noise source component of a breathy speech signal. The breathy noise source is characterized by a modulation that is a manifestation of gating by the vocal fold oscillations. The underlying model of generating breathiness is by no means complete – the physiology and aerodynamics of breathy speech is more complex and multi-faceted than our model represents.

Our preliminary analysis and pitch-modification of breathy vowels with the approach of this paper shows promise and warrants further research. This will include more accurate models of breathiness and its modification, refined decomposition methods, and a more complete analysis and synthesis of different forms of speech noise components for pitch-modification of running speech. In addition, benchmarking against more standard approaches and more formal listening evaluations will be performed.

7. ACKNOWLEDGEMENTS

Special thanks to MIT Lincoln Laboratory colleagues Mike Brandstein and Nick Malyska for insightful comments and discussion. We also thank Philip Jackson and Christine Shadle for providing the PSHF decomposition software that served as a reference for our simulation.

8. REFERENCES

[1] T. F. Quatieri and R. J. McAulay, "Shape-invariant time-scale and pitch modification of speech," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 40, pp. 497-510, 1992.

[2] X. Serra and J. Smith, "Spectral modeling synthesis: A sound analysis/synthesis system based on deterministic plus stochastic decomposition," *Computer Music Journal*, vol. 14, pp. 12-24, 1990.

[3] J. Laroche, Y. Stylianou, and E. Moulines, "HNS: Speech modification based on a harmonic + noise model," presented at IEEE International Conference on Acoustics, Speech, and Signal Processing, Minneapolis, MN, 1993.

[4] Y. Stylianou, J. Laroche, and E. Moulines, "High-quality speech modification based on a harmonic + noise model," presented at Eurospeech, 1995.

[5] P. J. B. Jackson and C. H. Shadle, "Pitch-scaled estimation of simultaneous voiced and turbulence-noise components in speech," *IEEE Transactions on Speech and Audio Processing*, vol. 9, pp. 713-726, 2001.

[6] B. Yegnanarayana, C. d'Alessandro, and V. Darsinos, "An iterative algorithm for decomposition of speech signals into periodic and aperiodic components," *IEEE Transactions on Speech and Audio Processing*, vol. 6, pp. 1-11, 1998.

[7] D. J. Hermes, "Synthesis of breathy vowels: Some research methods," *Speech Communication*, vol. 10, pp. 497-502, 1991.

[8] R. D. Kent and C. Read, *The Acoustic Analysis of Speech*. San Diego: Singular Publishing Group, Inc., 1992.

[9] K. N. Stevens, *Acoustic Phonetics*. Cambridge, MA: MIT Press, 1998.

[10] D. H. Klatt, "Software for a cascade-parallel formant synthesizer," *Journal of the Acoustical Society of America*, vol. 67, pp. 971-995, 1980.

[11] I. R. Titze, "Definitions and nomenclature related to voice quality," in *Vocal Fold Physiology: Voice Quality Control*, O. Fujimura and M. Hirano, Eds. San Diego: Singular Publishing Group, Inc., 1995, pp. 335-342.

[12] G. Fant, J. Liljencrants, and Q. Lin, "A four-parameter model of glottal flow," *Speech Transmission Lab. Quart. Prog. Status Rep.*, vol. 4, pp. 1-13, 1985.

[13] P. Boersma and D. Weenink, "Praat," http://www.praat.org

[14] Kay Elemetrics Corporation, "Disordered Voice Database," 1.03 ed, 1994.