Pitch-Scale Modification using the Modulated Aspiration Noise Source*

Daryush Mehta[1,2] and Thomas F. Quatieri[1,2]

[1]Speech and Hearing Bioscience and Technology
Harvard-MIT Division of Health Sciences and Technology
77 Massachusetts Avenue, Cambridge, Massachusetts, USA

[2]MIT Lincoln Laboratory
244 Wood Street
Lexington, Massachusetts, USA

[dmehta, quatieri]@ll.mit.edu

Abstract

Spectral harmonic/noise component analysis of spoken vowels shows evidence of noise modulations with peaks in the estimated noise source component synchronous with both the open phase of the periodic source and with time instants of glottal closure. Inspired by this observation of natural modulations and of fullband energy in the aspiration noise source, we develop an alternate approach to high-quality pitch-scale modification of continuous speech. Our strategy takes a dual processing approach, in which the harmonic and noise components of the speech signal are separately analyzed, modified, and re-synthesized. The periodic component is modified using standard modification techniques, and the noise component is handled by modifying characteristics of its source waveform. Since we have modeled an inherent coupling between the periodic and aspiration noise sources, the modification algorithm is designed to preserve the synchrony between temporal modulations of the two sources. The reconstructed modified signal is perceived in informal listening to be natural-sounding and typically reduces artifacts that occur in standard modification techniques.

Index Terms: pitch modification, aspiration noise, modulated noise, breathiness, voice quality

1. Introduction

The current study investigates the analysis and modification of aspiration noise in continuous speech, motivated in part by the need for better quality in concatenative speech synthesis applications. The approach builds on the linear source-filter modeling of speech [1] and research that aims at decomposing the speech signal into periodic and noise components for speech modification purposes [2, 3]. The source excitation is often modeled as an additive noise signal modulated at the pitch rate and synchronized with the voiced component before vocal tract filtering [4]. A challenge for analysis based on this model is accurate separation to estimate both temporal and spectral characteristics of the noise component that overlap with the periodic component. Previous researchers have documented the perceptual importance of noise modulations (e.g., [5]) and have further applied this understanding to the development of speech modification techniques [2], including our own work on pitch-scale modification of the breathy vowel [6].

*This work was sponsored by the Department of Defense 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.

This paper is organized as follows. Section 2 first describes selected approaches of current pitch modification algorithms and their limitations. Section 3 briefly reviews our earlier work on physiologically-based pitch modification of the breathy vowel, using a synchronous-modulation noise source model. Section 4 then presents our complete pitch-scale modification technique based on this model to process continuous speech. Finally, Section 5 presents an example of modifying continuous speech with the modulation-based approach, along with a preliminary comparative analysis with two standard algorithms.

2. Pitch-scale modification background

Several non-parametric methods have been developed for pitch modification [7]. In the popular technique of time-domain pitch-synchronous overlap-add (TD-PSOLA), the analysis time samples are set at instants of glottal closure that are estimated from the speech signal $s[n]$. Short-time frames of length N are centered on these instants, where N is an integer multiple of the local pitch period. The success of TD-PSOLA lies in its ability to smoothly duplicate or eliminate parts of the speech signal at a pitch-synchronous rate [7]. Drawbacks to the TD-PSOLA method include the generation of pseudo-periodicity of noise due to the replication of pitch periods and the requirement of accurate estimates of glottal closure time instants. The method also does not allow for separate control and modification of the noise signal, which can be achieved by parametric techniques that include our developed algorithm.

In parametric methods of pitch modification, the signal is fit to a particular model whose parameters are subsequently modified before signal reconstruction. In one parametric approach, speech modification systems are based on an analysis/synthesis system that models speech sounds as a sum of sinusoids [3, 8]. Even fricative sounds and plosive bursts are modeled using sinusoids. Each sinewave has a time-varying amplitude and time-varying phase associated with it:

$$s[n] = \sum_{k=1}^{M} A_k[n] \cos(\theta_k[n]), \tag{1}$$

where M is the number of sinewave components, $A_k[n]$ is the amplitude associated with the kth sinewave, and $\theta_k[n]$ is the phase of the kth sinewave. A degree-of-voicing measure sets a boundary frequency in the speech spectrum, below which voiced speech is assumed and above which noise is assumed. The sinewave frequencies themselves are chosen using a peak-picking algorithm in which the frequencies are not constrained

to be harmonically-related. The frequencies of the sinusoids in the voiced region are scaled by the desired modification factor while maintaining their spectral envelope, while the noise region is unmodified. Re-synthesis of the sinewaves completes the technique. Limitations include the possibility of inaccurate voicing measures and little control over the noise component at and between harmonic frequencies.

To handle noise more effectively, a modification algorithm based on a "harmonic + noise model" was developed [2]. The crux of this model is sinewave-based, where time-domain estimation is employed to determine amplitude, frequency, and phase parameters of sinusoids in a voiced region. An additional feature is separate modification of the noise component, which is assumed to be concentrated during the open phase of the pitch period. To account for this, a triangular envelope is imposed on a re-synthesized noise signal to result in the aspiration noise component. The envelope, however, is imposed *after* the noise source has been shaped by the vocal tract filter. In addition, in this and the basic sinewave-based approach, since aspiration noise has been shown to exist *across* the spectrum and not just at certain frequencies [9], a fullband decomposition technique would better estimate the noise component instead of assuming its energy is solely high frequency.

3. Modification of the breathy vowel

In previous work, we had developed a synchronous modulation noise-source model to aid in pitch modification of breathy vowels [6]. In this section, we review this approach and implementation as a basis for the design of a complete modification system on continuous speech.

3.1. Physiology of pitch control

One factor determining the fundamental frequency is vocal fold tension, which is dictated by properties of intrinsic muscles of the larynx. The stiffness of the body and cover of the vocal folds is largely due to the activity of the thyroarytenoid and cricothyroid muscles [10]. Since these muscles act more or less independently from changes in the vocal tract shape and since the glottal impedance is assumed to be large, we view the source mechanism as approximately decoupled from the filter.

Pitch-scale modification of speech signals can be performed by changing source excitation properties without affecting the spectral characteristics due to vocal tract resonances. Of particular interest is how the generation of turbulent noise is affected during a pitch change. The signal processing approach below assumes that modulations of the aspiration noise source follow the glottal waveform at the new fundamental frequency.

3.2. Pitch control model

Our model of pitch control is shown in Figure 1. According to the model, vowel production inherent assumes time-domain coupling between the periodic volume velocity waveform and the aspiration noise source. In this way, the model corresponds to the way that humans control the pitch of their voice. The source at Pitch 1 has a pitch period of T_0, while Pitch 2 represents a periodic source at a higher rate and thus having a shorter pitch period T_0'. Concomitantly, changes in the pitch period affect the function modulating the white aspiration noise source. Filtering mechanisms acting on the sources are assumed unchanged during pitch modification.

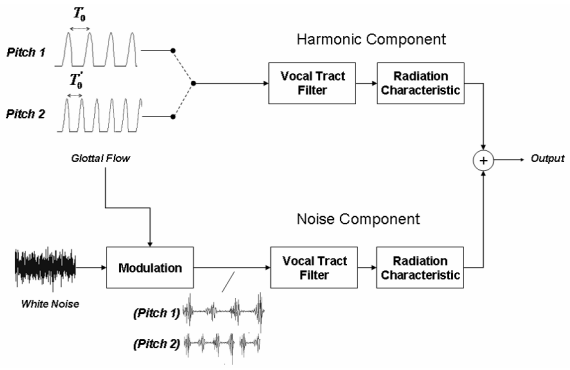

Figure 1 *Pitch control model of vowel with modulated aspiration noise source.*

3.3. Implementation of modifier

When performing pitch-scale modification, an algorithm should be able to recognize modulations present in the aspiration noise source component and preserve their synchrony with the periodic component. Our method of pitch-scale modification of vowels is based on reverse-engineering the pitch control model in Figure 1 to modify aspiration noise source characteristics. A schematic is shown in Figure 2.

The Decomposition block is a pitch-scaled harmonic filter technique [11], which separates the input into the harmonic and noise components $v[n]$ and $u[n]$, respectively. Pitch modification of the harmonic component can be accomplished by traditional algorithms (see Section 2) with the benefit of having the aspiration noise approximately removed.

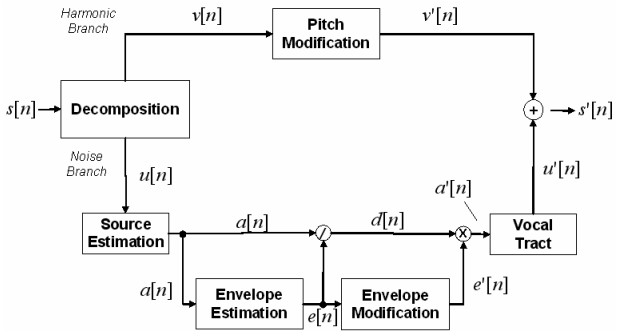

Figure 2 *Pitch-scale modification algorithm based on the pitch control model of Figure 1.*

The Source Estimation block estimates the underlying source waveform $a[n]$ through linear prediction analysis of the noise component to approximately remove the effects of the vocal tract filter and radiation characteristic. The modulation function $e[n]$ is then estimated so it may be modified and re-imposed on the demodulated excitation waveform $d[n]$. The Envelope Estimation block uncovers the noise modulations

generated by the glottal airflow fluctuations. We have chosen a method based on the Hilbert transform to perform estimation of the envelope $e[n]$ [6].

After uncovering the envelope of the noise source estimate, the algorithm re-modulates the glottal noise source $d[n]$ with the new envelope $e'[n]$, pitch-scaled by TD-PSOLA. The Spectral Coloring block re-imposes the spectral effects of the vocal tract and the radiation characteristic on the aspiration noise source using the parameters estimated in the Source Estimation stage. The modified signal $s'[n]$ is a sum of the modified harmonic and noise components, respectively $v'[n]$ and $u'[n]$.

4. Design of complete modification system

Extending our previous study on sustained vowel modification [6], we now process continuous speech and develop a system that handles both voiced and unvoiced speech by processing voiced speech with our system and then concatenating unprocessed unvoiced speech. The complete pitch-scale modification system is schematized in Figure 3.

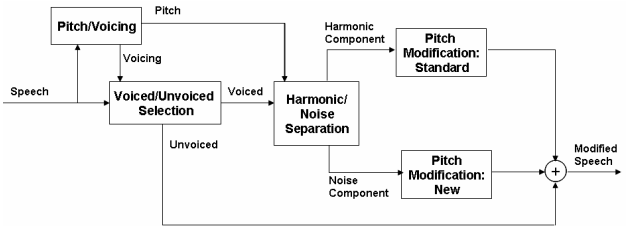

Figure 3 *Complete pitch-scale modification system for continuous speech. The noise component is processed by the lower branch of Figure 2.*

The previous algorithm in Figure 2 is amended to handle unvoiced speech sounds such as whispers, unvoiced fricatives, and silence regions. The new blocks are Pitch/Voicing and Voiced/Unvoiced Selection. Working on a frame-by-frame basis, the Pitch/Voicing stage determines the pitch of a short-time segment. Plausible pitch values label the segment voiced, and irrelevant values (i.e., close to zero) indicate an unvoiced or silence segment. Pitch estimation is accomplished using the speech signal processing tool Praat, which arrives at a periodicity measure by a forward cross-correlation analysis [12]. The pitch value is also needed in the Harmonic/Noise Separation stage since the harmonic filter technique is pitch dependent. Unvoiced speech sounds are additively combined with the modified harmonic and aspiration noise components to yield the new speech signal.

5. Modification example and observations on signal quality

Figure 4 displays spectrograms of the original and modified waveforms of continuous speech spoken by an adult female speaker: "As time goes by." The pitch scale is set to 1.75. A signal comparison is made between the output of our algorithm, the sinewave transformation system (STS) [3], and the TD-PSOLA modifier [7].

Using STS alone (Figure 4c), the noise component is perceived as somewhat tonal and perceptually separate from the periodic component. As seen around 1.2 s, aspiration noise between successive harmonics is undesirably reduced. Furthermore, harmonicity in the narrowband spectrogram near 0.7 and 1.1 s above 1500 Hz is overestimated by the STS algorithm. This latter effect is due to the degree-of-voicing index that selects a boundary frequency between periodic and stochastic spectral regions. Using TD-PSOLA alone (Figure 4d), discontinuities arise in transition regions, and noise can be distorted due to its periodic replication.

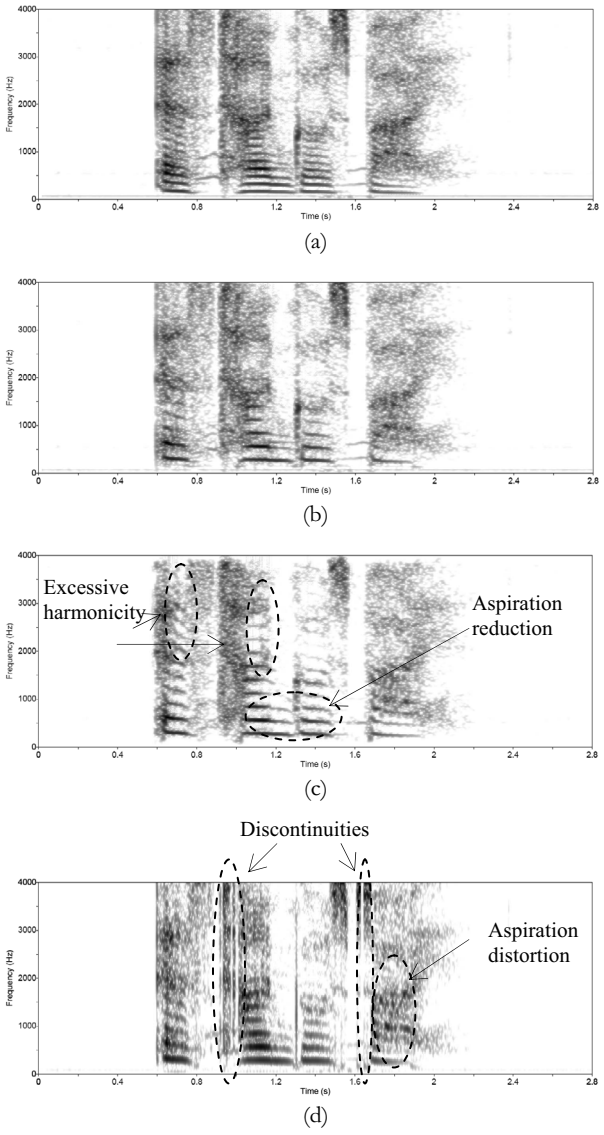

Figure 4 *Comparing pitch-scale modification of female utterance: "As time goes by." Pitch scale is 1.75. Spectrograms shown of (a) original signal, (b) modified by our algorithm, (c) modified by STS, and (d) modified by TD-PSOLA. Note (d) is generated with a shorter analysis window (16 ms versus 35 ms).*

Figure 4b shows the result of using our pitch-scale modification system of Figure 3, where STS is used as the harmonic modifier. By separating out the aspiration noise component, the conditions taxing STS are ameliorated, while the aspiration noise component is modified with a process tailored to its modulations (lower branch of Figure 2). The result is a modified signal with improved harmonicity and aspiration noise across the full speech spectrum. Likewise, we have found that when TD-PSOLA is used as the harmonic modifier (not shown), the result is a modified signal with reduced discontinuities and improved aspiration noise spectra.

In informal listening by a handful of experienced listeners, modified signals from the algorithm of Figure 3 are perceived to have a quality more consistent with that of the original waveform, compared with either of the two standard techniques alone (STS and TD-PSOLA). Specifically, the signal characteristics of the fullband aspiration noise appear to be better preserved, while the modified speech typically also has reduced artifacts that can occur in standard modification techniques, as was illustrated in the example of Figure 4.

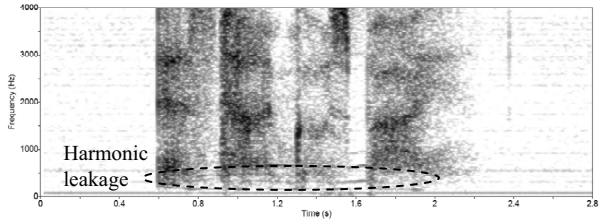

Figure 5 *Separated noise component of example in Figure 4.*

When performing harmonic/analysis on continuous speech, however, it is observed that the decomposed noise estimate sometimes contains leakage from the harmonic components. A minor case of this distortion is seen in Figure 5, illustrating the separated noise component in the example of Figure 4, in the 0.6–1.6 s region over roughly the 0–500 Hz range. This harmonicity in the noise spectrum, also observed when decomposing isolated vowel waveforms, especially with time-varying pitch and/or spectra, can be perceptually significant. The time-varying nature of spoken vowels, in addition to the effects of jitter and shimmer, may contribute to the suboptimal performance of the separation technique because of pitch estimation with inadequate temporal resolution.

6. Conclusions and future directions

In processing continuous speech signals, our pitch-scale modification algorithm aims at capturing aspiration noise source characteristics and suffers less from artifacts that are commonly observed in two standard techniques, such as those due to voicing errors in a sinewave-based approach [3] and glottal-closure estimation errors in TD-PSOLA [7].

Ultimately, we are interested in judging aurally the signal quality of our approach against standard and current methods, including the "harmonic + noise" model [2]. Such an evaluation, including judgment of the naturalness of the aspiration component and how it is perceived to blend with its corresponding periodic component, requires a rigorous listening test that is beyond the scope of the current effort. Nevertheless, informal listening shows promise for our technique for a variety of pitch-scale factors for synthesized vowels with varying pitch contours, as well as for continuous speech.

Issues have been raised regarding harmonic leakage in the separated noise component. In addition, further improvements are desired for estimation and modification of the aspiration noise envelope, as well as more accurate voicing and pitch estimation [9].

7. Acknowledgements

Special thanks to MIT Lincoln Laboratory colleagues Mike Brandstein and Nick Malyska for insightful comments and discussion.

8. References

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

[2] Y. Stylianou, J. Laroche, and E. Moulines, "High-quality speech modification based on a harmonic + noise model," *Proceedings of EUROSPEECH*, 1995.

[3] T. F. Quatieri and R. J. McAulay, "Shape invariant time-scale and pitch modification of speech," *IEEE Transactions on Signal Processing*, vol. 40, no. 3, pp. 497-510, 1992.

[4] D. H. Klatt, "Software for a cascade-parallel formant synthesizer," *Journal of the Acoustical Society of America*, vol. 67, no. 3, pp. 971-995, 1980.

[5] D. J. Hermes, "Synthesis of breathy vowels - Some research methods," *Speech Communication*, vol. 10, no. 5-6, pp. 497-502, 1991.

[6] D. Mehta and T. F. Quatieri, "Synthesis, analysis, and pitch modification of the breathy vowel," *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, 2005.

[7] E. Moulines and J. Laroche, "Nonparametric techniques for pitch-scale and time-scale modification of speech," *Speech Communication*, vol. 16, no. 2, pp. 175-205, 1995.

[8] M. W. Macon and M. A. Clements, "Sinusoidal modeling and modification of unvoiced speech," *IEEE Transactions on Speech and Audio Processing*, vol. 5, no. 6, pp. 557-560, 1997.

[9] D. Mehta, "Aspiration Noise during Phonation: Synthesis, Analysis, and Pitch-Scale Modification," Master of Science thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA, 2006.

[10] N.C.V.S., "Tutorials -- Voice Production -- How humans control pitch." Accessed January 23, 2006. http://www.ncvs.org/ncvs/tutorials/voiceprod/tutorial/cover.html.

[11] 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, no. 7, pp. 713-726, 2001.

[12] "Praat," version 4.4.04. P. Boersma and D. Weenink.