Synthesizing Breathiness in Natural Speech with Sinusoidal Modelling

Brett Matthews, Raimo Bakis and Ellen Eide
IBM TJ Watson Research Center
Yorktown Heights, New York 10598
brett@ece.gatech.edu, {bakis, eide} @us.ibm.com

Abstract
This paper discusses recent work in synthesizing a breathy quality in pre-recorded speech, which has applications in voice morphing and concatenative TTS. Previous work has shown that the breathy quality in speech is characterized in part by the presence of random noise in the upper region of the spectrum [1]. The sinusoidal modelling representation of speech facilitates making high-quality modifications to speech signals as well as modifying regions of the spectrum independently. We use sinusoidal modelling, along with techniques borrowed from analog communication systems to simulate aspiration noise in wideband speech signals above some lower cutoff frequency. Specifically, we use techniques based on amplitude modulation (AM) and phase modulation (PM), with the harmonics from the sinusoidal model of speech as carriers and lowpass random noise as the message signal. Formal listening tests were conducted and listeners rated the synthesized effect as "breathy" more often than in natural non-breathy speech, but significantly less often than in naturally breathy speech.

Index Terms: speech modification, voice conversion, sinusoidal modelling, concatenative TTS.

1. Introduction
In addition to the intended message sent from source to receiver, speech signals can convey other information such as speaker identity, and emotional state. Vocal qualities in the speech signal can enrich and emphasize (and sometimes even contradict) the information in the message. Breathiness can be characterized in an acoustic-articulatory sense by the presence of aspiration noise in higher frequency regions of the spectrum of the speech signal. In a perceptual sense, breathy speech is an important cue in determining, fatigue, emotional state and the presence of certain pathological conditions [2].

Breathiness plays an important role in Text-to-Speech Synthesis (TTS) systems as well. The presence of breathiness makes synthetic speech sound more natural and is an important feature in many rule-based TTS systems. While concatenative TTS systems accomplish the most natural-sounding synthetic productions of speech among state-of-the-art TTS, the use of pre-recorded speech limits the ability of these systems to control the degree of breathiness.

As breathiness is important in the perception of naturalness in speech, it is also important in the perception of speaker identity. The ability to add breathiness to pre-recorded speech has the potential to improve the quality of voice conversion algorithms significantly when the "target" speaker has a characteristically breathy voice.

In this paper we present recent work in synthesizing the breathiness quality in natural, pre-recorded speech. We use a sinusoidal modelling framework and lowpass random noise to synthesize the effect of aspiration in the upper regions of the spectrum in voiced sounds. We borrow basic concepts from Analog Communication Systems, specifically Amplitude Modulation (AM) and Phase Modulation (PM), to synthesize breathiness in speech.

Previous studies have used randomization [3] and random noise [4] [1] to devoice regions of the spectrum, but only in the context of purely synthetic speech. Stylianou [5] used modulated noise in a sinusoidal modelling context but only to model non-harmonic components (including aspiration) of existing speech signals, not to enhance or modify them.

The rest of the paper is organized as follows: Section 2 provides a brief review of message-carrier analog communication systems. In Section 3 we discuss our techniques for synthesizing breathiness in speech. In Section 4 we discuss our procedures for experiment and evaluation. Section 5 discusses our results and finally conclusions and future work are given in Section 6.

2. Analog Communication Systems
Analog communication systems are too varied in concept, application and implementation to be treated fairly in this short section. We limit this brief review to discuss only the concept of message-carrier modulation, wherein some time-varying property (typically the amplitude) of an information-bearing signal called the message is systematically encoded in another signal called the carrier. We further limit this discussion to continuous-wave modulation systems [6], which are characterized by having a sinusoidal carrier wave. Amplitude Modulation (AM) and Frequency Modulation (FM) are used in the broadcast transmission of audio and video signals and are, as such, the most familiar examples of continuous-wave modulation systems. AM and FM, along with Phase Modulation (PM) are briefly discussed in the next section.

2.1. AM, FM and PM systems
Expressions for the AM, FM and PM systems for analog communication are given in equations 1, 2 and 3, respectively.

\[ x_{AM}(t) = A_c [1 + A_{msg} x_{msg}(t)] \cos \omega_c t \]  

\[ x_{FM}(t) = A_c \cos (\omega_c t + 2\pi A_{msg} \int_0^t x_{msg}(\lambda)d\lambda) \]  

\[ x_{PM}(t) = A_c \cos (\omega_c t + A_{msg} x_{msg}(t)) \]

In each case, \( x_{msg}(t) \) is the message signal, \( \omega_c \) is the frequency of the carrier wave, and \( A_{msg} \) is a constant multiplier called the modulation index. In AM systems, the amplitude of the carrier wave is modified with proportion to the amplitude of \( x_{msg} \). In PM and FM, the message \( x_{msg}(t) \) modulates the argument (or angle) of the carrier sinusoid rather than the amplitude. In all cases the message signal \( x_{msg} \) can be recovered with varying fidelity when an inverse operation (demodulation) is applied at the receiver.
3. Synthesizing breathiness in speech

In our analysis, we find that the breathiness quality is characterized in part by the presence of modulated noise in higher frequency regions of voiced speech. A female speaker was asked to say the same phrase twice, with and without a breathy quality in her voice. Recordings of the utterances were made at 22kHz sampling rate and their spectrograms (cut at 4kHz) are given in figure 1. A noise-like presence above 1.5kHz is clearly seen in the spectrogram of the breathy version of the utterance in figure 1 (b), but not in the normal, or non-breathy version in figure 1 (a). Most notably we consistently found that harmonics in the upper region of the spectrum are easily distinguishable in perceptually non-breathy speech, but not in breathy speech. The rest of this section discusses our methods for synthesizing this effect in prerecorded speech.

3.1. Sinusoidal modelling framework

Sinusoidal modelling [7] has become a popular framework for making high quality modifications to speech signals. In addition, the sinusoidal modelling representation of speech also allows for modifying regions of the spectrum independently. For these reasons we use a modified version of the system in [8]; specifically, we use an experimental version of its reconstruction/synthesis phase. Equation 4 describes our system for reconstructing the $l^{th}$ voiced frame of speech

$$s^l(t) = \sum_{k=0}^{N(l)} A^l_k(t) \cos \left( k\omega_c t + \phi^l_k \right)$$

where $N(l)$ is the number of harmonics in frame $l$ and $\phi^l_k$ is the phase at the time analysis instant for the frame. $A^l_k(t)$ is the time-varying amplitude function for the $l^{th}$ voiced frame and is determined by interpolating amplitude values $A^l_k$ and $A^{l+1}_k$ at time analysis instants for the $l^{th}$ and $(l+1)^{st}$ voiced frames. The time-varying function for the fundamental frequency $\omega_0(t)$ is similarly determined from $\omega_0^l$ and $\omega_0^{l+1}$. A DFT representation is used for unvoiced frames, which are synthesized with a simple inverse DFT operation.

3.2. Message/Carrier model

Although there are many important differences in the frequency characteristics of linear (AM) and exponential (FM and PM) continuous-wave modulation systems, the transmitted waveform in each case is a bandpass signal centered at the carrier frequency $\omega_c$. With respect to an unmodulated carrier sinusoid, the transmitted waveform is significantly more spread out in frequency. We exploit this property of continuous-wave modulation systems to synthesize the characteristic of breathiness in speech. Specifically, we use the harmonics in our sinusoidal modelling framework as carriers and lowpass-filtered random noise as the message in voiced frames. Unvoiced frames are not modified.

As previously discussed, we characterize breathy speech by the presence of random noise at higher frequency regions of the spectrum. We create a noise function $x_{msg}(t)$ and use some modulation technique to apply it to regions of the spectrum above a specified frequency threshold $\omega_{co}$. $x_{msg}(t)$ is created by generating 100 frames (5 ms) of uniform random noise and applying to it a lowpass elliptical IIR filter with a cutoff frequency $\omega_{BW}$. The signal is then normalized to have unity standard deviation and used as the message in the AM and PM systems as described in the following sections.

3.3. PM in sinusoidal modelling

In our system for synthesizing breathiness, phase modulation is applied only to voiced frames according to equation 5.

$$s_{PM}^l(t) = \sum_{k=0}^{N(l)} A^l_k(t) \cos \left( k\omega_0 t + \phi^l_k + \alpha_{msg} x_{msg}(t) \right)$$

The message signal $x_{msg}(t)$ is scaled by a constant $\alpha_{msg}$ and simply added to the argument of the oscillator (i.e. the cosine function) from equation 4. Modifications are made only to harmonics with frequency $k\omega_0$ such that $k\omega_0 > \omega_{co}$.

Three parameters in our PM system are easily varied to modify the degree of breathiness: the message amplitude $\alpha_{msg}$, a lower cutoff frequency for speech modification $\omega_{co}$, and the bandwidth of the message signal.

Spectrograms (cut at 4kHz) of an utterance from a female speaker before and after phase modulation is applied, are given in figures 2 (a) and 2 (b), respectively. The lower cutoff frequency for speech modification $\omega_{co}$ is 1kHz. It is clear from the figure that harmonics of the speech signal with frequencies greater than $\omega_{co}$ have been modified. The bandwidth $\omega_{BW}$ of the message signal in this example is 100Hz.

Since the frequency characteristics of FM and PM are similar, only PM was used in our study.

3.4. AM in sinusoidal modelling

Amplitude modulation (AM) systems predate FM and PM and are used when low fidelity communication is adequate. In the context of broadcasting an information-bearing signal through the airwaves, FM and PM are superior to AM. We show in this section that some of the same properties which disadvantage AM in communication systems may make it more favorable for synthesizing breathiness with sinusoidal modelling.

Since the spectrum of an AM signal is essentially a frequency-translated version of the message signal’s spectrum, the bandwidth is always roughly twice that of the message signal. The spectrum of an FM or PM signal is more complex and its bandwidth is not so easily predicted given the message signal $x_{msg}(t)$. For this reason the effects of synthesizing breathiness on the spectrum of the speech signal are more easily controlled and analyzed when AM-based techniques are used instead of PM.

Two methods for synthesizing breathiness based on AM were developed in this study, and are given in equations 6 and 7.

\[\text{Equation 6}\]

\[\text{Equation 7}\]

For a detailed analysis of the acoustic correlates of breathy speech, the reader is referred to [1].
The previous section describes two frameworks (AM and PM) for speech modification, each of which has 3 important parameters: the lower cutoff frequency for modification $\omega_{co}$, and the bandwidth and amplitude of the message signal, $\omega_{BW}$ and $\alpha_{msg}$. In all cases, higher values of $\omega_{BW}$ and $\alpha_{msg}$ increase the effect of the modification on the speech signal, but can also cause displeasing distortion if set too high. The lower cutoff frequency $\omega_{co}$ controls the region of the spectrum to be modified. Setting $\omega_{co}$ too low ($< 1kHz$) also introduces displeasing artifacts into the speech signal.

In our experiments we sought to synthesize breathiness with minimal degradation in the audio quality of the speech. For this reason we chose conservative values of $\omega_{co}$, $\omega_{BW}$ and $\alpha_{msg}$. In preliminary tests (with aggressive parameters) we found that the pitch-synchronous multiplier term in equation 7 caused a noticeable degradation in audio quality, and was not used in our experiment.

4. Experimental procedure

The previous section describes two frameworks (AM and PM) for speech modification, each of which has 3 important parameters: the lower cutoff frequency for modification $\omega_{co}$, and the bandwidth and amplitude of the message signal, $\omega_{BW}$ and $\alpha_{msg}$. In all cases, higher values of $\omega_{BW}$ and $\alpha_{msg}$ increase the effect of the modification on the speech signal, but can also cause displeasing distortion if set too high. The lower cutoff frequency $\omega_{co}$ controls the region of the spectrum to be modified. Setting $\omega_{co}$ too low ($< 1kHz$) also introduces displeasing artifacts into the speech signal.

In our experiments we sought to synthesize breathiness with minimal degradation in the audio quality of the speech. For this reason we chose conservative values of $\omega_{co}$, $\omega_{BW}$ and $\alpha_{msg}$. In preliminary tests (with aggressive parameters) we found that the pitch-synchronous multiplier term in equation 7 caused a noticeable degradation in audio quality, and was not used in our experiment.

4.1. Listening tests

To evaluate our system we conducted formal listening tests. 28 native speakers of North American English, 15 female and 13 male, were recruited to listen to 24 sentences and rate them according to two criteria: audio quality and breathiness. The subjects were first asked to rate the audio quality of each sentence they heard on a scale of 1 (poor) to 5 (excellent). They were then asked whether they would characterize each sentence as “breathy” or not; only responses of “yes” and “no” were accepted for this question. The sentences were played in random order, and listeners were not told that any synthetic modifications had been made to the speech signal. To account for the effects of list order, one half of the listeners were given a playlist of sentences in reverse order.

We chose one set of parameters for AM and PM and applied them to 12 speech utterances. We chose the AM algorithm, with
Table 1: Experimental procedure. A breakdown of the conditions under which sentences in the listening test were recorded and modified and the number of sentences in each category. The AM and PM algorithms were only applied to “normal” sentences.

<table>
<thead>
<tr>
<th>System</th>
<th>Original</th>
<th>Unmodified</th>
<th>AM</th>
<th>PM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>3</td>
<td>3</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>Instructed Breathy</td>
<td>3</td>
<td>3</td>
<td>6</td>
<td>6</td>
</tr>
</tbody>
</table>

Table 2: Mean opinion scores (MOS) of audio quality.

<table>
<thead>
<tr>
<th>System</th>
<th>MOS overall</th>
<th>Unmodified overall</th>
<th>AM</th>
<th>PM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>3.78</td>
<td>3.36</td>
<td>3.53</td>
<td>3.25</td>
</tr>
</tbody>
</table>

Table 3: Breathiness Ratings. The fraction of listeners, for each sentence category, who responded “yes” to the breathiness question.

6.67% of sentences in the normal group. 22.67% of PM sentences and 15.33% of AM sentences were rated as breathy by the listening group. While both AM and PM rate higher than normal, listeners overall rated them much lower than “instructed breathy.”

6. Conclusions and future work

The initial results show that, to some listeners, the breathy quality was effectively synthesized, but that there is room for improvement. Future plans include increasing the degree of breathiness in contexts where it is usually found, e.g. in vowels following consonants, especially plosives and fricatives. We also intend to make dynamic changes to the lower cutoff frequency for modification, particularly so that it corresponds to the location of the third formant [1].

7. References


5. Results

Results from formal listening tests are summarized in tables 2 and 3. Table 2 gives mean opinion scores (MOS) of audio quality for the “original” and “unmodified” sentences along with those modified by the AM and PM algorithms. Listeners rated the baseline sinusoidal modelling system, represented by the “unmodified” category, lower than “original” by 0.42. AM rates higher than PM by 0.28. Surprisingly, sentences to which the AM algorithm were applied rate close to original (a difference of 0.25) and actually rate higher than “unmodified”. The likely cause of this particular result is a mismatch in the recording environment, i.e. in some of the recordings a low but observable noise (the hum of a desktop PC) is present.

Breathiness ratings are given in table 3. As expected, ratings were low for the normal group of sentences, and high for “instructed breathy.” Specifically, 84.67% of “Instructed Breathy” sentences were rated as breathy by the listening group, along with 6.67% of sentences in the normal group, 22.67% of PM sentences and 15.33% of AM sentences were rated as breathy by the listening group. While both AM and PM rate higher than normal, listeners overall rated them much lower than “instructed breathy.”