# AUTOMATIC DYSPHONIA RECOGNITION USING BIOLOGICALLY-INSPIRED AMPLITUDE-MODULATION FEATURES

*Nicolas Malyska, Thomas F. Quatieri, and Douglas Sturim*

MIT Lincoln Laboratory
{nmalyska, quatieri, sturim}@ll.mit.edu

**ABSTRACT**

A dysphonia, or disorder of the mechanisms of phonation in the larynx, can create time-varying amplitude fluctuations in the voice. A model for band-dependent analysis of this amplitude modulation (AM) phenomenon in dysphonic speech is developed from a traditional communications engineering perspective. This perspective challenges current dysphonia analysis methods that analyze AM in the time-domain signal. An automatic dysphonia recognition system is designed to exploit AM in voice using a biologically-inspired model of the inferior colliculus. This system, built upon a Gaussian-mixture-model (GMM) classification backend, recognizes the presence of dysphonia in the voice signal. Recognition experiments using data obtained from the Kay Elemetrics Voice Disorders Database suggest that the system provides complementary information to state-of-the-art mel-cepstral features. We present dysphonia recognition as an approach to developing features that capture glottal source differences in normal speech.

## 1. INTRODUCTION

The ability to recognize characteristic voice qualities is an intriguing human trait. With this ability, we can obtain information such as a speaker's identity, state of health, and degree of fatigue. The acoustic properties that convey these elements are continually being understood. Our research is motivated by the desire to develop features that capture a class of *source* mechanism characteristics related to voice qualities.

Automatic recognition systems in speech technology often focus on representing the vocal tract of speakers, but the source properties tend not to be explicitly analyzed. One area of automatic recognition that has only recently begun to emerge is automatic speech disorder, or *dysphonia*, recognition. A dysphonia is a disorder of the voice production mechanisms in the larynx with specific perceptual, acoustic, and physical correlates. Examples of these disorders include excessive tension of the laryngeal muscles and the presence of abnormal masses of tissue on the vocal folds [16]. The problem of dysphonia recognition is particularly interesting because it is largely dependent on differences in the glottal source, rather than differences in the vocal tract resonances. Dysphonic speech also may represent the *extremes* of acoustic phenomena occurring in normal voices such as the irregular nature of glottalization [12]. Because dysphonia recognition deals specifically with differences in the voice source mechanisms, there is hope that the techniques found in this domain can be applied to other recognition problems such as speaker recognition.

This paper provides evidence that the dysphonic voice can be modeled as frequency-band-dependent amplitude fluctuations. We relate these fluctuations to communications engineering concepts in order to build an AM synthesis model of dysphonic voice. A biologically-motivated analysis model is then explored with which to capture these modulations. We present evidence from a set of GMM-based dysphonia recognition experiments that our model captures complementary voice information to state-of-the-art mel-cepstral features.

***This work is sponsored by the United States Air Force Research Laboratory 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.**

## 2. AMPLITUDE MODULATION MODEL FOR VOICE

### 2.1. Envelope Fluctuations in Voice

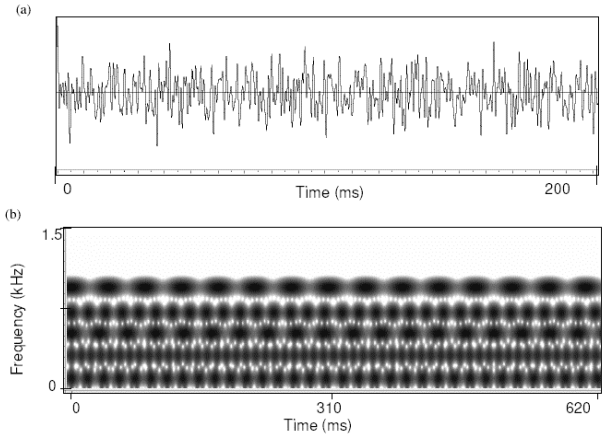


**Figure 1.** A synthesized sum of amplitude-modulated sinusoids illustrating the concept of bandwise envelope fluctuations. The time waveform (a) shows little periodicity, whereas there is clear structure in the spectral domain (b).

An interesting source characteristic common in the dysphonic voice is time variation of spectral amplitude envelopes. These fluctuations can occur in different bands and interact to produce complicated time-domain behavior. One way to illustrate this

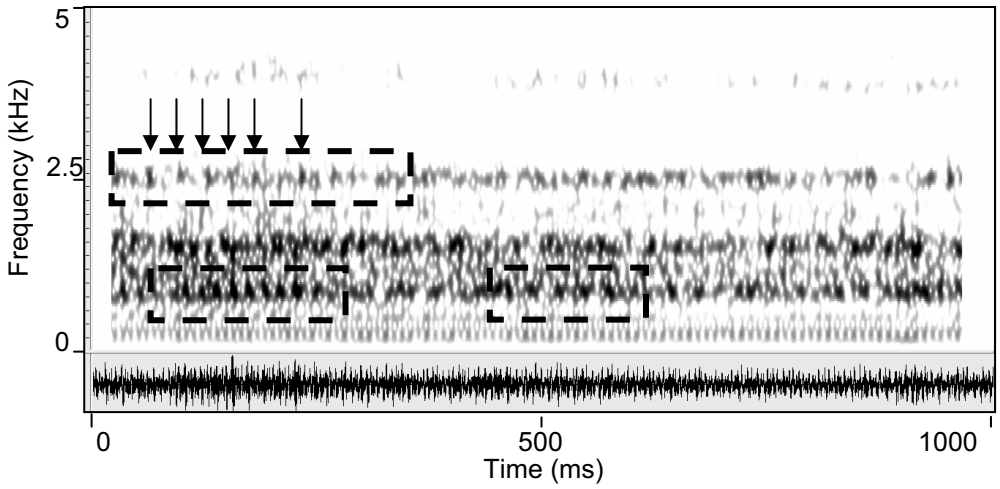


**Figure 2.** Spectrogram of a sustained vowel produced by a female voice patient. A 10-ms Hamming window was used in generating this figure. Top left box: 30-to-40-Hz pattern around F3; peaks highlighted by arrows. Bottom left box: 50-to-60-Hz pattern around F1. Right box: AM at about 30 Hz near F1. Different bands show different AM patterns.

phenomenon is shown in Figure 1, where a sum of sinusoids is depicted, each amplitude-modulated by a different sinusoidal envelope. Whereas the sum of the signals yields a time waveform without clear periodicity or structure, the spectrogram reveals clear amplitude-modulated envelopes.

An example of these bandwise fluctuating patterns in real dysphonic voice taken from the Kay Elemetrics Voice Disorders Database [1] is shown in Figure 2. This image depicts the spectrogram of a female voice patient producing the sustained vowel /a/. We notice that throughout the spectrogram, there are different regions of repeating patterns; the boxes in the figure highlight prominent examples of these patterns that are different in different frequency bands.

Other observations of fluctuating patterns in dysphonic speech have been made [8] including repeating patterns of glottal pulse amplitudes, high-frequency pulses with sinusoidal amplitude envelopes, and wideband noise with repeating amplitude patterns [5]. Our work suggests that many fluctuating spectral envelopes seen in dysphonia can be modeled by sinusoidal carriers, each modulated by a different set of sideband components. In some cases the sinusoidal carriers are the harmonic line components of the glottal waveform. In the frequency domain, the sidebands are visible as smaller neighboring components [8]. As we will show in the next section, this view is consistent with an engineering definition of AM.

### 2.2. AM Model

Band-dependent envelope fluctuations are not unique to the voice, one example being their occurrence in the field of communications engineering. Consistent with our observations of dysphonia, we propose an AM model for dysphonic voice whereby a series of *bandlimited* signals are transposed to higher frequencies. This process amounts to modulating the amplitude envelopes of a series of sinusoidal carriers. In mathematical form, amplitude modulation of a single sinusoid using a bandlimited envelope is defined in [2] as:

$$s(t) = g(t)\cos(\omega_c t)$$

where

$$g(t) = A_c[1 + m(t)]$$

is the envelope multiplied by the cosine carrier with radian frequency $\omega_c$ to create the modulated signal $s(t)$. The bandlimited source signal, $m(t)$, is used to create the envelope by adding it to unity and scaling the sum by a constant $A_c$. Here $m(t)$ is defined to be between -1 and 1 such that $g(t)$ always remains positive. Also, $\omega_c$ is assumed to be no less than twice the highest frequency component of the original signal.

Analysis in our model is performed much in the same way that demodulation is done in the communications domain, as shown in Figure 3. First, frequencies around the carrier are isolated by bandpass filtering, and the result is passed through an envelope detection stage. In our work, we use incoherent detection using the envelope of the analytic signal obtained by the Hilbert transform.

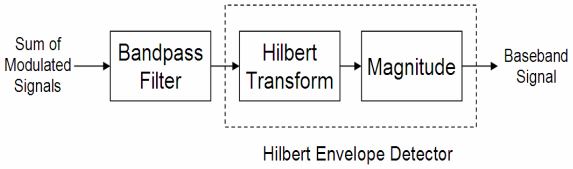


**Figure 3.** Model for the analysis of amplitude modulations in voice using bandpass filtering followed by envelope extraction using the magnitude of the analytic signal.

Thus, our model presents the human voice as created by a series of summed amplitude-modulated sinusoidal sources with non-overlapping bandwidths. Researchers including Teager and Titze have highlighted the importance of understanding the speech signal as a glottal source carrier modulated by physiological inputs such as muscle movements, vortices of air, and the motion of laryngeal tissues. [14, 15]. The human voice, however, is almost surely not well modeled solely in this way. The physical sources of sound produced by the glottis and by turbulent airflow in the vocal tract are for the most part not well described as single sinusoids. Another issue is that there is also

probably an overlap between the carriers and sidebands of physiologically plausible speech sources. Nevertheless, our representation is consistent with our observations of real dysphonic speech and may provide the basis for more complex models.

### 2.3. Biologically-Inspired AM Analysis

There is evidence that structures for AM analysis as described above exist in the human brain in an area called the inferior colliculus (ICC). In particular, the ICC is thought to extract the frequency content of the modulation envelopes applied to different frequency bands of auditory stimuli. By modeling the ICC after a design introduced by Dau, Kollmeier, and Kohlrausch [3, 7], we can construct a biologically-inspired modulation feature extraction system.

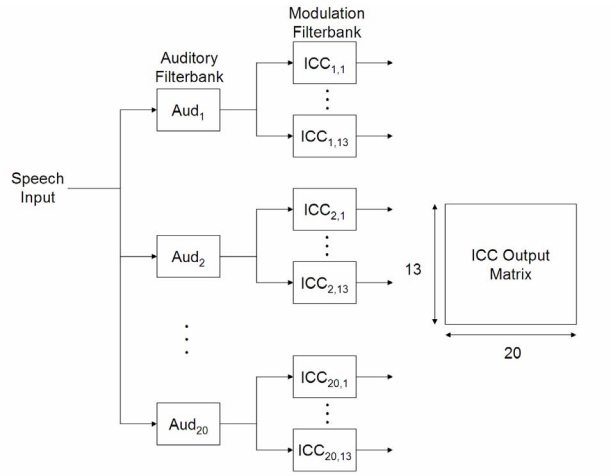


**Figure 4.** Architecture of the auditory model including the inferior colliculus modulation filtering stage.

The first stage of the ICC model processes the input signal with a bank of 20 mel-spaced cochlea-like gammatone filters. Following this stage, the envelope of each of these output channels is filtered by a second bank of 13 modulation filters with exponentially spaced center frequencies from 12 to 107 Hz. The envelopes resulting from this process yield a 13-by-20 element output matrix. The sum is then taken across the columns yielding a vector of 13 elements per frame. This process is shown schematically in Figure 4 (further details can be found in [8, 11]).

## 3. AUTOMATIC DYSPHONIA RECOGNITION

We devised an automatic system to categorize speech as either pathological or normal. By using the above auditory model sensitive to bandwise modulations, we hypothesized that we would be able to capture information unique to dysphonia, improving performance of the system over one using mel-cepstral features alone.

### 3.1. Methods

The voice corpus used for these experiments was the Kay Voice Disorders Database [1]. For our tests, we used the 12-second continuous Rainbow Passage utterances. Each file was antialias-filtered and downsampled to a 8-kHz sampling rate, the lowest common denominator of the dataset.

The Kay voice database is split into two groups—normal and pathological. Due to issues with the contents of the database discussed in [8], a number of pathological utterances were discarded from our evaluation. This yielded 397 pathological and 53 normal voices which were used in a normal/pathological recognition experiment. A jackknife method was devised to produce 5 different groupings of the voices, each using approximately 80% of the utterances for training and 20% for testing. In this way, all of the utterances could be used for testing without the problem of overlapping training and testing sets.

To extract features, each utterance was first run through the ICC auditory model of Figure 4 at a 5-ms frame interval or the standard mel-cepstrum using 20 filters at a 10-ms frame interval. The cepstrum of each model-output vector was computed, discarding the zeroth bin. RASTA and cepstral-mean subtraction channel-compensation techniques [6] were then applied and delta features computed for each feature set.

The features were then processed by a GMM-based pattern classifier [13]. Two models, one for normal voice and one for pathological utterances, were trained and the test utterances were compared against them. Each test yielded a likelihood score which was used to create a detection-error tradeoff (DET) curve to judge performance, giving false-alarm versus miss probability for different score thresholds.

### 3.2. Results

DET curve results for standard mel-cepstrum, ICC, and the linear fusion of the two techniques using an exhaustive search method is plotted in Figure 5 [8]. As shown, the equal-error rate (EER)—the point where the false-alarm probability equals the miss probability—of mel-cepstrum alone is 3.77 percent, with ICC yielding 5.66 percent EER, and a linear fusion of the scores resulting in 2.02 percent EER. Both the ICC model and the fusion improve performance over mel-cepstrum, with the fusion resulting in a benefit over the entire DET curve. This result gives evidence that the ICC model provides complementary dysphonia information to the mel-cepstrum.

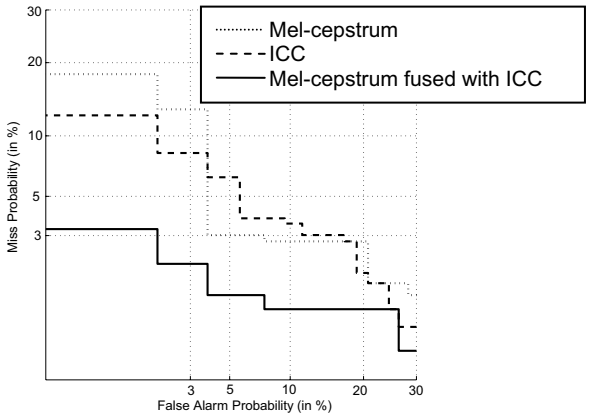


**Figure 5.** Performance of the mel-cepstrum compared with the modulation features.

## 4. DISCUSSION

Objective techniques exist in the speech pathology literature by which to measure time fluctuating amplitude envelopes in the voice. Most popular approaches are based on measures of either glottal-pulse timing and amplitude perturbations such as jitter, shimmer or the harmonic-to-noise ratio. These measures work reliably only on steady vowel portions of speech and are sensitive to estimates of pitch [9, 10]. Exactly how our AM model relates to these measures is not known, although both are expected to capture some of the same amplitude envelope phenomena.

Dysphonia recognition using the Kay database is reported in the literature by several groups. Dibazar and Narayanan [4] describe a normal/pathological classification system consisting of a GMM classifier and using mel-cepstral features. Using a somewhat different subset of the normal and pathological Rainbow Passage they report a 2.54 percent EER [4]. Objective clinical perturbation measures, such as shimmer and jitter can also be used as a model for feature extraction. The best previous attempts using this technique on the Kay database sentences obtained 95.6 percent recognition for the Rainbow passage [10]. Our results suggest improved performance when compared to these studies.

It is also possible to perform recognition experiments for individual voice disorders such as paralysis and lesions on the vocal folds. Our research has included the implementation of several versions of specific-dysphonia recognition experiments, but this problem has proved significantly more difficult than the normal/pathological recognition task. A vocal-fold paralysis recognition system, for example, performed at 30 percent equal-error-rate level in the best case [8]. There has not been extensive reporting on specific dysphonia recognition in the literature although [4] discusses one series of experiments with a throat muscle-tension disorder called A-P squeezing.

One potential problem with the Kay database is that some of the normal speakers were recorded at different sites and over potentially different channels than the pathological voices. To test the possibility of having developed a channel recognition system rather than a dysphonia recognition system, we ran a preliminary experiment with only the non-speech portions of the utterances used in the normal/pathological distinction. An energy-based speech detection software program, a standard part of the Lincoln Laboratory GMM system, was used to select non-speech frames. As the energy threshold for speech detection was lowered, the DET curves became worse, eventually exhibiting about 25 percent EER. Therefore, although further study of the database properties is needed, there is evidence that suggests that our system primarily detects dysphonic speech and not the channel on which it was recorded.

## 5. CONCLUSIONS

We have presented a model by which the voice can be represented as a series of summed amplitude-modulated sinusoids. This view is motivated by observations of dysphonic voice that provide evidence of band-dependent amplitude modulations in the human voice. A biologically-inspired model was built to analyze modulations in speech and dysphonia recognition by using this new model. Results with this technique showed improvements over standard mel-cepstral features, especially when a linear fusion was implemented. This research supports the view that a bandwise amplitude modulation analysis provides complementary information about the voice source. Although we have focused on extreme voice types in dysphonic speech, the underlying technology may be useful in more general speech technology applications.

## 6. REFERENCES


[1] "Voice Disorders Database," Version 1.03 ed: Kay Elemetrics Corp., 1994.

[2] L. W. Couch, *Digital and Analog Communication Systems*, 5th ed. Upper Saddle River, NJ: Prentice Hall, 1997.

[3] T. Dau, B. Kollmeier, and A. Kohlrausch, "Modeling auditory processing of amplitude modulation .1. Detection and masking with narrow-band carriers," *Journal of the Acoustical Society of America*, vol. 102, pp. 2892-2905, 1997.

[4] A. A. Dibazar and S. Narayanan, "A system for automatic detection of pathological speech," presented at Conference Signals, Systems, and Computers, Asilomar, CA, 2002.

[5] B. R. Gerratt and J. Kreiman, "Toward a taxonomy of nonmodal phonation," *Journal of Phonetics*, vol. 29, pp. 365-381, 2001.

[6] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Transactions on Speech and Audio Processing*, vol. 2, pp. 578-589, 1994.

[7] G. Langner and C. E. Schreiner, "Periodicity coding in the inferior colliculus of the cat .1. Neuronal mechanisms," *Journal of Neurophysiology*, vol. 60, pp. 1799-1822, 1988.

[8] N. Malyska, "Automatic voice disorder recognition using acoustic amplitude modulation features," MS Thesis in Department of EECS. Cambridge, MA: MIT, 2004.

[9] P. Milenkovic, "Least mean-square measures of voice perturbation," *Journal of Speech and Hearing Research*, vol. 30, pp. 529-538, 1987.

[10] V. Parsa and D. G. Jamieson, "Acoustic discrimination of pathological voice: Sustained vowels versus continuous speech," *Journal of Speech Language and Hearing Research*, vol. 44, pp. 327-339, 2001.

[11] T. F. Quatieri, N. Malyska, and D. Sturim, "Auditory signal processing as a basis for speaker recognition," presented at IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, Mohonk Mountain, NY, 2003.

[12] L. Redi and S. Shattuck-Hufnagel, "Variation in the realization of glottalization in normal speakers," *Journal of Phonetics*, vol. 29, pp. 407-429, 2001.

[13] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, pp. 19-41, 2000.

[14] H. M. Teager and S. M. Teager, "Evidence for nonlinear sound production mechanisms in the vocal tract," in *Speech Production and Speech Modeling*, vol. 55, *NATO Adv. Study Inst. Series D*, H. W.J. and A. Marchal, Eds. Bonas, France: Kluwer Academic Publishers, 1990, pp. 241-262.

[15] I. R. Titze, "Workshop on acoustic voice analysis. Summary statement," National Center for Voice and Speech, Denver, CO 1995.

[16] W. R. Wilson, J. B. Nadol, Jr., and G. W. Randolph, *Clinical Handbook of Ear, Nose, and Throat disorders*. New York, NY: The Parthenon Publishing Group, 2002.