DOI: 10.17469/O2104AISV000001

KHALIL ISKAROUS

The encoding of vowel features in Mel-Frequency Cepstral Coefficients

Most work on acoustic phonetics uses formant frequencies as the parameterization of the phonetic signal for understanding the acoustic difference between the sounds of the world's languages. Work in speech technology, however, has relied for several decades on Linear Prediction Coefficients (LPC) and Mel-Frequency Cepstral Coefficients (MFCC's), due to their greater invariance to physical differences between speakers. This paper explores the phonetics of the MFCC's, asking whether these coefficients can be used by phoneticians to develop a greater understanding of the phonetic nature of speech segments. This is done through an analysis of the ability of individual coefficients to distinguish between American English vowels in the Hillenbrand database.

Keywords: Mel-Frequency Cepstral Coefficients, vowel features.

1. *Introduction*

For the last seventy years, the most popular parametrization of the speech signal in the field of Linguistic Phonetics has been the frequency spectrum (linear, bark, or mel-tranformed), whether parameterized in terms of its formants or moments (e.g., Potter, Kopp & Green, 1947; Joos, 1948; Odden, 1991; Boersma, Escudero & Hayes, 2003; Forrest, Weismer, Milenkovic & Dougall, 1988; Labov, 1994). The spectrum has been used to characterize vowel and consonant inventories across languages of the world, sociophonetic differences, and the influence of prosody on segmental production. In contrast, the field of speech technology started to move away from the spectrum and its formant peaks about 50 years ago for the main reason that the shape of the spectrum varies enormously across speakers, especially when children are included. This is for the simple reason that the spectrum, even when bark or mel-transformed, is highly sensitive to vocal tract length which can vary from 5-10 cm in children to 15-18 cm in adults. Early work in speech coding established Linear Prediction Coefficients (LPC) and Reflection Coefficients (RC), methods based on multiple and partial regression analysis (Wakita, 1973), as parameterizations of the speech signal that are robust to speaker variability. Later work identified the closely related Mel-Frequency Cepstral Coefficients (MFCCs) (Mermelstein, 1976) to be especially robust to large variation in speakers, and became the most popular parametrization for speech recognition, since the mid 1980's. Despite a few exceptions (e.g. discrete cosine coefficients, Harrington & Cassidy, 1999), the fields of Linguistic Phonetics and Speech Technology have proceeded largely in parallel over the last several decades, each using its separate characterization of speech acoustics. This situation is perplexing, since speaker-independence should not be a concern only for the speech technologist, but also for linguists interested in the acoustic distinctions across dialects and across prosodic domains within a dialect, *regardless of the specific physical characteristics of the speaker*. A possible reason for this parallel procession of highly inter-related fields is that since Joos (1948), we have known, at least for vowels, how to phonetically characterize the spectrum: the Front/Back contrast is characterized largely by F2, the High/Low Contrast is characterized largely by F1, and the Round/Unround contrast is characterized by the lowering of both F1 and F2. In contrast, little, if anything, is known about how vowel features are encoded in the MFCC coefficients. The goal of this paper is to initiate an understanding of the phonetics of MFCC's, and specifically how the basic vowel features of Front/Back, High/Low, and Round/Unround are encoded in these coefficients.

One hypothesis is that the vowel features are encoded in a highly distributed way across the MFCC's, so that no single coefficient codes for a single vowel feature. The competing hypothesis is that the vowel features are encoded in specific coefficients. Which of these possibilities holds makes the project of understanding the phonetics of MFCC's of possible interest to speech technologists, not just to linguists. The reason is that a chief motivation in the field of speech technology is to express the information in the speech signal that is speaker-independent with as few bits as possible. So, if the linguists have been correct in the hypothesis that the differences between vowel segments are specifiable with very few pieces of information, the feature settings (e.g. Front/Back, High/Low, and Round/Unround), and if it is the case that individual MFCC's code these vowel contrasts, then vowels could be specified with a smaller number of coefficients, those that specify the featural contrasts. That is, we would establish a hierarchy of importance amongst the MFCC's. Therefore, investigating the phonetics of MFCCs is possibly of interest to linguists and technologists. After all, these two groups are basically interested in the same thing, efficient encoding of speech, whether it's for the purpose of efficiently describing linguistic systems or for enabling efficient technologies. The overall aim of this program, therefore, is to bridge a gap between two fields that have a common interest in speech invariants and speech variation, which started five decades ago. The paper will concentrate on the Front/Back, High/Low, and Round/Unround features only, and will leave features such as Tense/Lax, Nasal/Oral, and the consonant features for future work. Also left for future research are time-varying effects, such as coarticulation.

2. *MFCCs*

The aim of this section is to highlight the meaning of MFCC's. One aspect is the mel-frequency transformation, emphasizing low frequencies, and averaging across higher frequencies, and is already quite familiar to linguists as it is a common transformation of the spectrum (Ladefoged, 1996). Since formant peaks for vowels are quite narrow in bandwidth, emphasizing the low frequencies, allows for higher resolution necessary for vowel identification. High Frequency burst peaks for stops and sibilant noise have higher bandwidth, therefore averaging across large spans of higher frequency still allows for consonant identification. The Cepstral aspect of MFCC's part is less familiar, except in the particular application to F0 extraction, which will not be discussed in this paper. Cepstral Coefficients have a long history in the signal processing and geophysics literatures (Robinson, 1954; Bogert, Healy & Turkey, 1963), which was reviewed recently in Oppenheim, Schafer (2004).

Before presenting the formula for how to compute MFCC's, it is useful to understand *why* they were invented. The title of the first paper to present Cepstral coefficients in detail, Bogert et al. (1963), is quite telling "The quefrency analysis of time-series for echoes". The purpose of the cepstrum was to find echos in time series. In geophysics, where they were developed, the interest is in seismogram time series, where signal reflections from the earth are processed to find echos from significant structures like oil or earthquakes (Silvia, Robinson, 1978). The idea of the cepstrum is to reveal important information about *where* the crucial echoing structures are in the medium from which the signal emerges, by processing the signal in a particular way (to be discussed momentarily). If we regard the vocal tract as an echoing chamber where glottal and supra-laryngeal sound signals are reflected in constrictions and the glottal and lip ends of the vocal tract (Wakita, 1972), with the speech signal emerging at the lips after all these echoing events, we could see how the same cepstral technique could be useful for revealing useful information about the constrictions in the vocal tract, potentially revealing information about Vowel Constriction Location and Degree, which are basically equivalent to vowel features like Front/Back, High/Low, and Round/Unround. The basic finding of Bogert et al. (1963) is that the crucial information about the echo generating structures can be found from taking the Fourier Transform of the (Log of) the Fourier Transform (i.e., the spectrum that has been at the center of acoustic phonetics). The reason why this works is based on the idea of homomorphic signal processing, and was soon recognized by Schafer and Oppenheim (Schafer, 1968; Oppenheim, 1964). The basic intuition can be seen in two steps: 1) the source-filter theory explains the spectrum of vowels as the multiplication of the transfer function of the vocal tract filter and the spectrum of the glottal source in the frequency domain, since the independent variables of the functions are frequency; 2) taking the Fourier Transform of the Log of the Fourier Transform yields a replacement of the spectrum by an *addition*, instead of a multiplication, of the cepstrum of the vocal tract filter and the cepstrum of the glottal wave. Since the spectrum of the glottal wave is rich in high frequencies (harmonics) it yields a spike in higher (quefrency) of the cepstrum, and can be *liftered* out, leaving the low quefrency vocal tract contribution. Since the technique is closely related to the traditional spectrum, with its amplitude and phase, Bogert et al. (1963) used a word game to develop names for the cepstral quantities: the *spectrum* became the *cepstrum*, the independent variable of *frequency* became *quefrency*, the *magnitude* of the spectrum became the *gamnitude,* the phase became *saphe,* and filtering became *liftering*. The algorithm for measuring the MFCC's is quite simple:

a. Obtain the Fourier spectrum of a portion of the speech signal (e.g. 25 ms).
b. Mel-transform the frequency scale, emphasizing low frequencies.
c. Obtain the logarithm of the mel-transformed Fourier spectrum.
d. Obtain the Fourier spectrum of the previous result.
e. Lifter out the effect of the glottal wave.

Due to the ubiquity of MFCC's in speech technologies, many computer languages have libraries for computing them. In this paper, the Python python_speech_features library was used.

3. *Methods*

An ideal data set for this investigation is the Hillenbrand vowel database (Hillenbrand, Getty, Clark & Wheeler, 1995). This data set contains utterances by 45 men, 48 women, 27 boys, and 21 girls producing hVd for all the American English vowels. The large number of speakers, the inclusion of different ages, and the near staticness of the vowel in the h_d context, allows us to understand the phonetics of MFCC's for static vowels in data with high speaker variability, but little temporal variability. The vowels were classified by the author, uncontroversially, for their settings of their basic vowel features (Table 1). As mentioned earlier the features Tense/Lax which could distinguish [i] and [ɪ] for instance, are not being investigated in this paper.

Table 1 - *Vowels of American English and their feature specifications*

	Front/back	*High/Low*	*Round/Unround*
i	Front	High	Unround
ɪ	Front	High	Unround
e	Front	NA	Unround
ɛ	Front	Low	Unround
æ	Front	Low	Unround
ə	NA	NA	NA
u	Back	High	Round
ʊ	Back	High	Round
o	Back	NA	Round
ɔ	Back	Low	Round
ɑ	Back	Low	Unround
ɚ	NA	NA	NA

The database contains the beginning, steady state, and end of each vowel. 25 ms were extracted from middle of each steady state, a hamming window was applied, and the resulting speech waveform was preemphasized. The python_speech_features *mfcc* function was applied, yielding 13 MFCC's. The results for two sampling rates will be presented:16,000 Hz and 8,000 Hz. The results to be presented are sensitive to these specific choices of sampling rates, a problem that will be discussed later in Section 5. There were a total of 1668 measurements (540 men vowels, 576 women vowels, 324 boy vowels, 228 girl vowels).

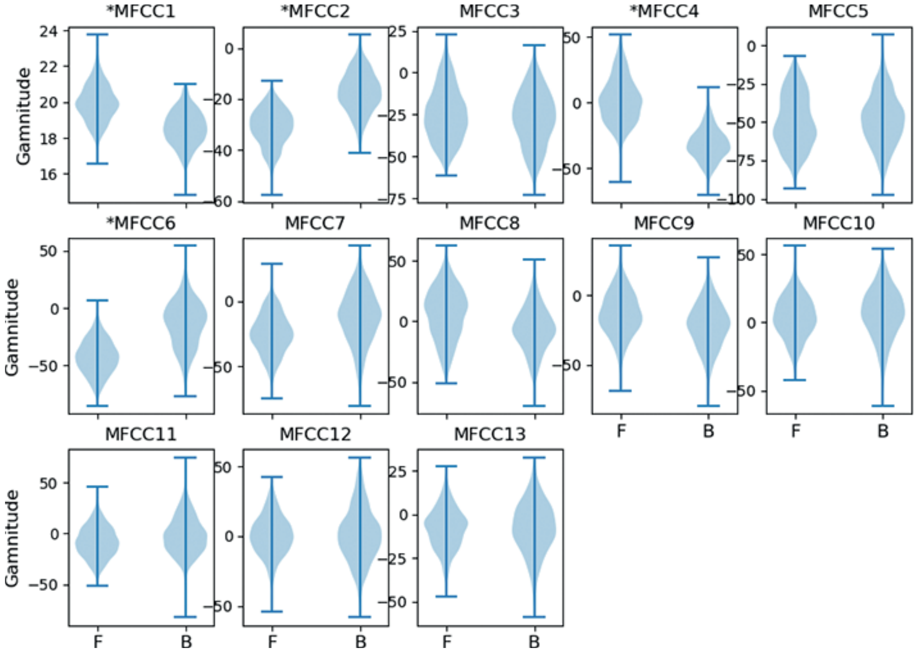
4. *Results*

The goal of this work is to understand how each of the basic vowel features is encoded in the MFCCs. To achieve this goal, for each feature, we compared the vowels specified oppositely for that feature on each of the MFCC's. The results for 16,000 Hz will be discussed before the results for 8,000 Hz. Figure 1 shows the results for the Front/Back feature. Each panel shows how Front and Back vowels compare on each of the MFCC's by showing the distribution of that coefficient's data for all the speakers. To evaluate whether each of the coefficients captures the Front/Back distinction, an ANOVA was run for each of the coefficients, and the effect was deemed significant, if p < .001, a low significance level, due to Bonferroni correction for 39 (3 Features x 13 coefficients) tests. In addition, to directly test the distance between the distributions for each coefficient, the Cohen's d effect size measure was used for measuring the distance between distributions A and B:

$$\text{Cohen's d }(A, B) = \frac{\text{Mean}(A) - \text{Mean}(B)}{\sqrt{\frac{\text{sd}(A)^2 + \text{sd}(B)^2}{2}}}$$

Cohen's d estimates the distance between means of the distributions in units of standard deviations. In Figure 1, significance of a distributional contrast is indicated by an asterisk before the word MFCC at the top of each panel. Significance was determined through both p < .001 for the ANOVA test and Cohen's d being larger than 1 standard deviation in magnitude.

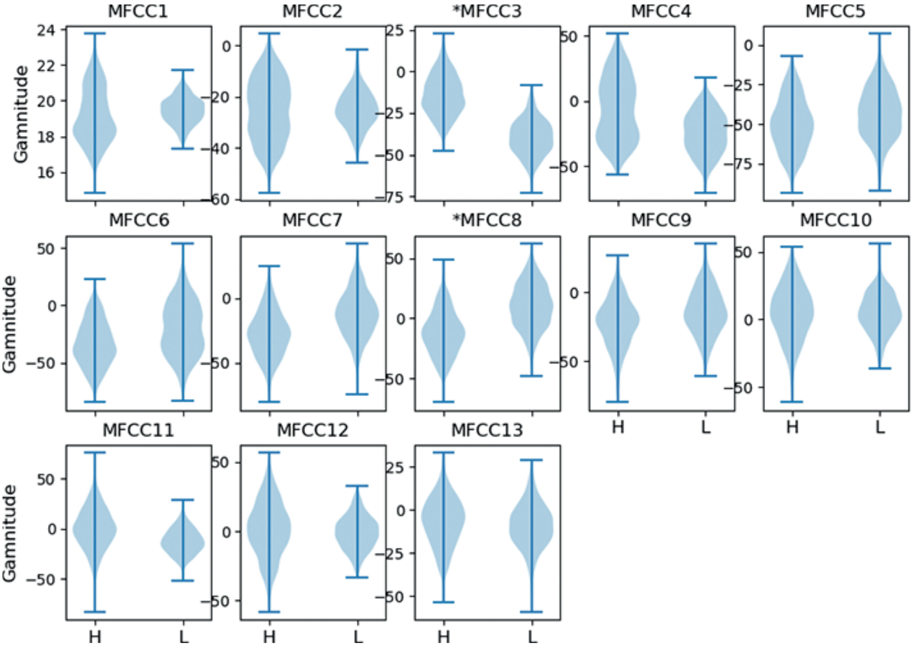
Figure 1 - *Front/Back encoding by MFCC's*



As can be seen from Figure 1, MFCC's 1, 2, 4, and 6 *directly* and significantly distinguish between Front and Back vowels. This does not mean that the other MFCC's do not code for vowel frontness, since they could do so in complex combinations with each other. But it does mean that at least one of the MFCC's, indeed 4, directly encode one of the most important features used by linguists to describe vowel systems.

Figure 2 shows the analogous situation for vowel height. As can be seen, MFCC 3 and 8 *directly* code for the height contrast. Two important things emerge from this result. One is that the MFCC's for distinguishing Front/Back (1, 2, 4, 6) do not overlap with those that distinguish High from Low (3, 8). This conforms to the idea from linguistic phonetics that the features describe different aspects of a segment (Jakobson, Fant & Halle, 1952).

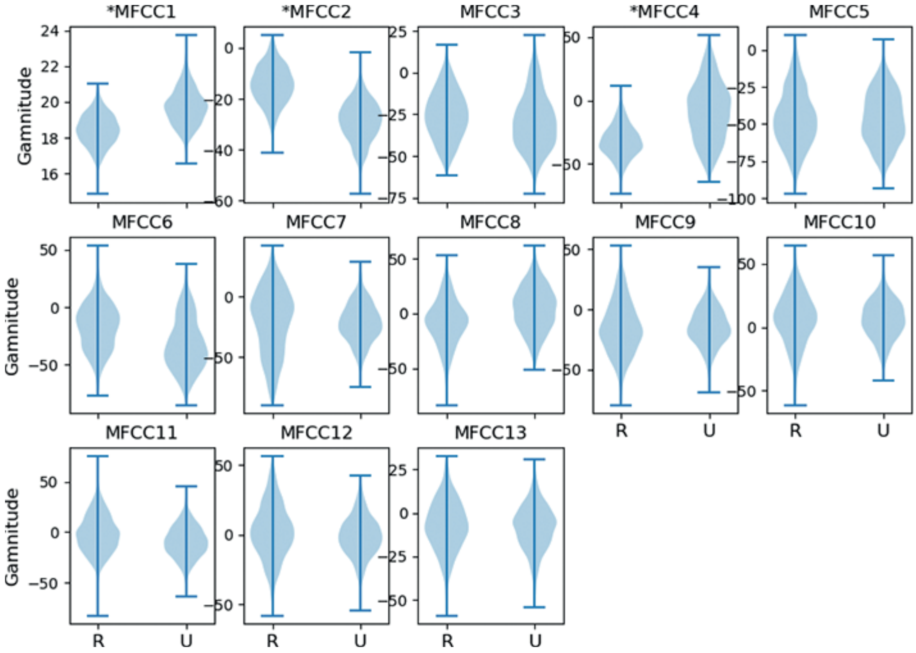
Figure 2 - *High/Low encoding by MFCC's*



The other thing that emerges is that the number of MFCC's that directly encode Front/Back is significantly larger than the number of MFCC's encoding High/Low significantly. There are several possible reasons for this. One is an inheritance from the spectrum, in which F1, the spectral indicator of Height, ranges over a smaller frequency range (approximately 200-1000 Hz) than F2 (approximately 800-2800 Hz), the spectral indicator of Frontness. Since the cepstrum is based on the spectrum, one would expect major aspects of spectral structure to affect cepstral structure. Another possibility is that if we relax the significance criteria, more MFCC's would show significant results, therefore the difference would not be indicative of anything substantial about the phonetic encoding of MFCC's.

Figure 3 presents the results for the Round/Unround feature.

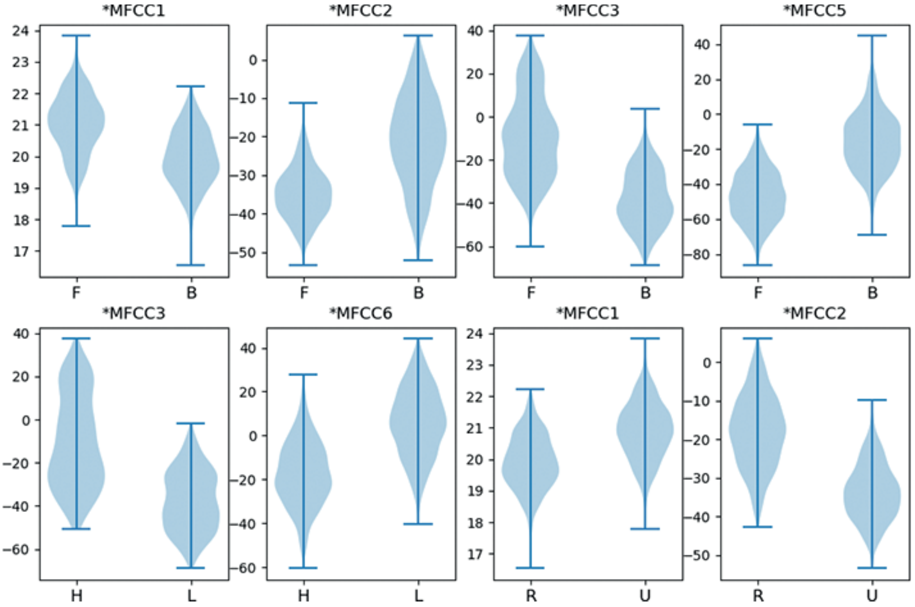
Figure 3 - *Round/Unround encoding by MFCC's*



The major thing that emerges is that the direct indicators of Round/Unround are a subset of those for Front/Back, something that phoneticians have usually refer to for English as "backness predicts rounding". Note also that the direction of the effect supports the linguistic fact that Front vowels and Unrounded vowels are similar, whereas Back vowels and Rounded vowels are similar. To determine if MFCC's can directly encode rounding in a non-overlapped manner with frontness, one must use datasets for languages that have front rounded vowels (like French) or back unrounded vowels (like Japanese), which is left for future research.

Figure 4 presents the significant coefficients for the same dataset and features, when the sampling frequency was halved at 8,000 Hz. Front/Back is encoded via MFCC's 1, 2, 3, 5. High/Low is encoded via MFCC's 3 and 6. Round/Unround is encoded via MFCC's 1 and 2. Therefore at this sampling rate, there is indeed overlap between the encoding of Front/Back and High/Low in MFCC 3. However, MFCC 6 encodes High/Low, and not Front/Back. Further investigation will need to take place to reveal why the same MFCC significantly encodes two linguistic distinctions. The other result emerging from Figure 4 is that Rounding is redundant, as was seen with the 16,000 Hz data.

Figure 4 - *Significant encodings of vowel features for 8,000 Hz sampling rate*



5. *Discussion and conclusion*

In this paper, it's been shown that there are *individual* MFCC's that code the linguistically motivated vowel features used for many decades. The main drawback of this work is that we get different results for different sampling rates, which is something unfamiliar to linguists investigating speech acoustics, where statements about F1 and F2 of the spectrum are made regardless of the sampling rate used. What is even more problematic is that for the higher sampling rate, 16,000 Hz, we had no overlap in the MFCC's encoding Front/Back and High/Low, whereas at the lower sampling rate, 8,000 Hz, we get some partial overlap.

The first issue, having to qualify statements about acoustic phonetics with information about the sampling rate used, is not as problematic as it may seem, however. The reason that there is sensitivity to sampling rate in MFCC, as well as other parameterizations such as Linear Prediction Coefficients and Reflection Coefficients, is that these coefficients encode information about the reflectivity of acoustic waves in the vocal tract, and sampling rate is directly related to the assumed number of tubes we assume the vocal tract to consist of (Wakita, 1972), and inversely related to the assumed average length of the vocal tract through the formula: *SamplingRate* = Mc/2l, where *M* is the number of tubes, *c* is the speed of sound in air, and *l* is the assumed average length of the vocal tract. The higher the sampling rate, the higher the assumed number of tubes. Therefore, it should not be surprising that the information in these coefficients is sensitive to sampling rate.

The question then arises as to which sampling rate to use. Familiarity with the spectrum for many years has taught us that if we care only about vowels, a sampling rate of about 8,000 Hz is sufficient, since the first three formants are below 4,000 Hz. However, if we care about consonants as well, and there is no general reason to not care about them, then 16,000 to 20,000 Hz is sufficient, since the distinctions between sibilants for instance can be detected between 5,000 and 10,000 Hz. If we sample at higher rates, e.g. the current default of 44,100 Hz, it is quite easy to downsample to these rates. Therefore, it is possible to pick an optimal sampling rate like 16,000 or 20,000 Hz, and make acoustic phonetic statements, assuming whichever of them the field settles on. The use of a higher sampling rate like 16,000 Hz also seems to solve the other problem we saw, which is overlap between the phonetic encodings of Front/Back and High/Low, a problem which needs further investigation since 8,000 Hz is sufficient for the investigation of vowels, which is what was being done in this paper.

Another issue that arises is redundancy. All the features seem to be encoded by more than one MFCC. This is somewhat surprising, since MFCC's are usually assumed to be highly uncorrelated, but it is not really a problem, since one can use multiple indicators of the same feature, leading to more robust classifications. Further investigation of how Tense/Lax is encoded, and whether the encoding of this feature overlaps with the other features, should reveal how the MFCC's separately and jointly represent linguistic distinctions.

Bibliography

Boersma, P., Escudero, P. & Hayes, R. (2003). Learning abstract phonological from auditory phonetic categories: An integrated model for the acquisition of language-specific sound categories. *Proceedings of the 15th International Congress of Phonetic Sciences*, 1013-1016.

Bogert, B.P., Healy, M.J.R. & Turkey, J.W. (1963). The quefrency analysis of time series for echoes. In Rosenblatt, M. (Ed.), *Proc. Symp. Time Series Analysis*. New York: Wiley, 209-243.

Forrest, K., Weismer, G., Milenkovic, P. & Dougall, R. (1988). Statistical analysis of word initial voiceless obstruents: Preliminary data. In *The Journal of the Acoustical Society of America*, 84, 115-123.

Harrington, J., Cassidy, S. (1999). *Techniques in Speech Acoustics*. Dordrecht: Kluwer Academic.

Hillenbrand, J., Getty, L.A., Clark, M.J. & Wheeler, K. (1995). Acoustic characteristics of American English vowels. In *The Journal of the Acoustical Society of America*, 97, 3099-3111.

Jakobson, R., Fant, G. & Halle, M. (1952). *Preliminaries to speech analysis: The distinctive features and their correlates*. Technical Report 13. Massachusetts: Acoustics Laboratory, MIT.

Joos, M. (1948). *Acoustic Phonetics*. Issue 23 of Language Monographs, Linguistic Society of America Language Supplement.

Labov, W. (1994). *Principles of Linguistic Change, Volume 1: Internal Factors*. Wiley-Blackwell.

Ladefoged, P. (1996). *Elements of Acoustic Phonetics.* The University of Chicago Press.

Mermelstein, P. (1976). Distance measures for speech recognition, psychological and instrumental. *Proceedings of the Joint Workshop on Pattern Recognition and Artificial Intelligence*, Held at Hyannis, Massachusetts, June 1-3, 1976.

Odden, D. (1991). Vowel geometry. In *Phonology,* 8, 261-289.

Oppenheim, A. (1964). *Superposition in a class of nonlinear systems*. Ph.D. Dissertation, MIT.

Oppenheim, A.K., Schafer, R.W. (2004). From frequency to quefrency: a history of the cepstrum. *IEEE Signal Processing Magazine*, 21, 95-106.

Potter, R., Kopp, G. & Green, H. (1947). *Visible Speech.* New York: Van Nostrand.

Robinson, E.A., (1954). *Predictive Decomposition of Time Series with Applications to Seismic Exploration*. Ph.D. Dissertation, MIT.

Schafer, R. (1968). *Echo removal by discrete generalized linear filtering*. Ph.D. dissertation, MIT.

Silvia, M., Robinson, E. (1978). Use of the kepstrum in signal analysis. In *Geoexploration,* 16, 55-78.

Ulrych, T.J. (1971). Application of homomorphic deconvolution to seismology. In *Geophysics*, 36(4), 650-660.

Wakita, H. (1973). Direct estimation of the vocal tract shape by inverse filtering of acoustic speech waveforms. In *IEEE Transactions on Audio and Electroacoustics*, 21, 417-427.