A DESCRIPTION AND COMPARISON OF THE FEATURE SETS USED IN SPEECH PROCESSING

Richard Duncan
Mississippi State University
Mississippi State, Mississippi 39762, USA
e-mail: duncan@isip.msstate.edu
Ph (601) 325-3149 - Fax (601) 325-3149

ABSTRACT — There are two primary recognition tasks in speech processing, speech recognition and speaker recognition. From analyzing a sound wave, characteristics of the speaker's articulatory organs can be estimated, these characteristics offer a means for biometric identification and authentication. Speech recognition systems seek to understand the content of the spoken waveform. Most current research in recognition technology is towards speaker-independent systems, which can convert the speech of any speaker. While these two goals seem completely orthogonal, they both pull deeply from the same pool of signal processing algorithms for feature extraction. The challenge in both fields is to find a set of features robust to variations in the environment. This paper provides an overview of the feature extraction algorithms used for both speech and speaker recognition. It concludes with a brief evaluation of the different signal modeling algorithms discussed with small recognition experiments.

1. INTRODUCTION

The aim of speaker recognition is to recognize from the speaker's voice characteristics of the speaker's articulatory organs and speaking manner for identification purposes. The structure of the vocal tract, the size of the nasal cavity, and vocal chord characteristics can all be estimated through signal analysis [7]. Speaker recognition is the broad term applied to both speaker identification and verification. For identification the estimated speaker characteristics are compared against a database of registered users for the closest match. For verification the identity claim of a speaker is accepted or rejected based on his or her biometric signature [11].

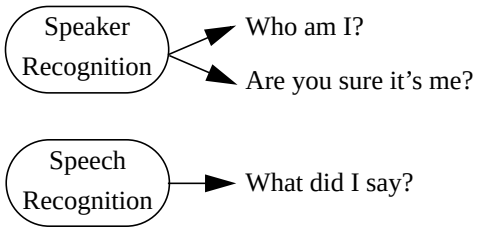

Figure 1. Different Tasks

Speech recognition seeks to translate a spoken acoustic signal into words. Humans make words by moving the articulators through a predictable series of positions. If these sequences can be extracted from the signal then the spoken words can be recognized. Many applications for speech recognition call for speaker-independent systems; these products try to recognize the speech of any speaker.

Even though these two goals seem completely orthogonal, they both perform pattern recognition on speech data. Some systems, such as the Nuance 6 server, perform both speech recognition and speaker verification simultaneously [9]. Due to this similarity of procedure both technologies suffer from the same pitfall, a large performance degradation arises due to environmental differences between training and testing conditions [11]. In short, the performance of the technology is closely tied to the environment in which the systems are developed, so noisy real-world situations lead to sub-optimal performance.

The algorithms used by speech processing front-ends are based on acoustic models of the vocal tract and ear canal [2]. The next section will motivate the need for feature extraction with a brief overview of the pattern recognition. This will be followed by a description of the prevalent front-end algorithms in use today. The final section discusses channel normalization techniques and the associated implications on speech and speaker recognition.

2. PATTERN RECOGNITION

A pattern recognition system contains two components: A front-end feature extractor and a classifier. The hope is that once data is transformed into the feature space data from the same class will be pulled as close together as possible and data from different classes will be pushed as far apart as possible. Once a classifier is trained to distinguish between classes in this transformed feature space, a recognition system need only transform input data through the same feature

extraction system and determine into which class a new observation falls [7].

There are two significant problems with applying this approach to speech processing. First, there is an implicit assumption that the training and testing conditions are comparable. The use of a different microphone, background noise, and transmission channels can cause drastic performance degradation — a key criterion for judging the robustness of a feature set is its invariance to such channel variation [7]. Secondly, there is significant overlap of classes in the feature space. Zhao has prepared plots to demonstrate this overlap in two corpora of speech data collected over the telephone network [16]. Speech recognition engines attempt to overcome this overlap problem by using powerful statistical processing to incorporate language modeling. Such techniques are beyond the scope of this paper, but Deshmukh, Ganapathiraju, and Picone present an overview of a hierarchical speech recognition engine [3].

3. SIGNAL MODELING ALGORITHMS

The purpose of signal modeling (commonly referred to as feature extraction) is to transform audio data into a space where observations from the same class will be grouped together and observations of different classes will be pushed apart [7]. These transformations are chosen based on physiological studies of the human auditory and articulatory systems. For example, articulators cannot move from one position to another in less than about five milliseconds, so practical systems can sample the spectrum 100 times a second with only a minor loss in resolution [12].

Speech is a dynamic signal, so we are interested in examining the short-term spectrum. The frame duration is defined as the length of time over which a set of parameters is valid. While frames are not overlapping, we typically use an overlapping analysis window to draw in a larger number of speech samples for each spectral measurement. Performing direct spectral analysis on such a small amount of data is equivalent to applying a sharp rectangular window to a signal, which produces spectral distortion. The frequency response of a rectangular pulse is a sinc function, $(\sin x)/x$, which has a curved pass band and a large amount of ripple in the stop band. Different window shapes are realized by applying a weighting function. The Hamming window,

$$w(n) = \frac{\alpha_w - (1-\alpha_w)\cos(2\pi n/(N_s-1))}{\beta_w}, \quad (1)$$

is a specific instance of the Hanning window with $\alpha_w = 0.54$. The β_w parameter is chosen for normalization so that the energy of the signal will be unchanged through the operation. The Hamming window's shape provides spectral analysis with a flatter pass band and significantly less stop band ripple, both properties are important for obtaining smoothly varying parametric estimates. Most state-of-the-art systems today use a frame duration of 10 milliseconds and a window duration of 25 milliseconds [8].

One feature extracted from the signal is the absolute energy of the spectrum. The other category of spectral measurement is the energy at specific frequencies. These measurements are similar to the initial stages of transduction in the human auditory system — the hair cells in the cochlea serve a similar purpose. There are three ways to obtain these spectral measurements: Direct application of a digital filter bank in the time domain, using the Fourier Transform, and linear predictive analysis [12]. The later two methods are more prevalent in today's systems due to computational efficiency.

Since human hearing is not equally sensitive across a linear scale, we map the spectrum onto a perceptual frequency scale. Experiments in human perception have shown that frequencies within a certain bandwidth of a nominal frequency cannot be individually distinguished, this is known as a critical bandwidth. The Bark scale is one perceptual scale in which frequencies are mapped to critical band rates [12]. The mel scale is a simpler approximation that maps the perceived pitch of a tone onto a linear scale. Stevens and Volkman empirically determined a mapping between the mel scale and real frequencies in 1940. The scale is roughly linear below 1000 Hz and logarithmic above 1000 Hz [12].

Simple Fourier transform-based filter banks designed for front-ends obtain the desired frequency resolution on a mel-scale. To implement this filter bank, the window of speech data is transformed into the frequency domain by the Fourier Transform. Once in the frequency domain each filter bank amplitude coefficient can be found through the application of a linear combination of the spectrum and the frequency response of the desired filter. In practice overlapping triangular filter banks are used where the center frequency of one filter serves as the endpoint of its two adjacent filters (Young 1996). Thus, each filter bank amplitude coefficient represents the average spectral magnitude in the filter channel,

$$S_{avg}(f) = \frac{1}{N_s}\sum_{n=0}^{N_s} w_{FB}(n)|S(f)|, \quad (2)$$

where N_s represents the number of samples used to obtain

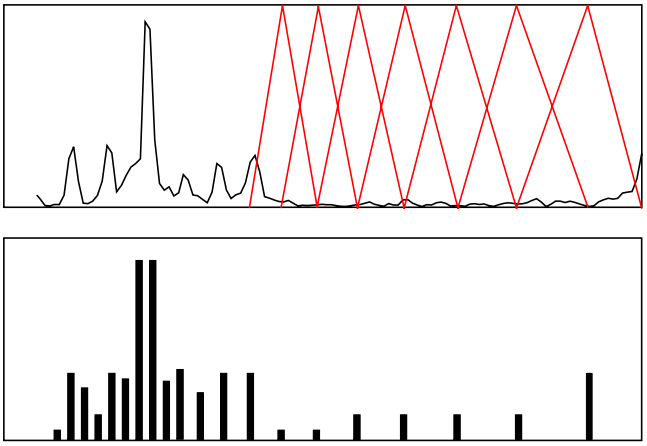

Figure 2. Mel-spaced triangular filter banks

the averaged value, $w_{FB}(n)$ represents the weighting function (a triangle function is described above), and $S(f)$ is the magnitude of the frequency response computed by the Fourier Transform [12].

Linear predictive (LP) analysis is a means for obtaining the smoothed spectral envelope of $P(w)$ through an all-pole model of the power spectrum. The linear predictor coefficients have a direct correlation to log-area ratios, the geometric parameters of the lossless tube model for speech production [1]. Filter bank amplitudes are obtained by sampling the LP spectral model at the appropriate filter bank frequencies. This can be done through a direct evaluation of the LP model, but in practice the Fourier Transform is applied to the predictor coefficients. Since there are fewer LP coefficients than audio samples this approach is more computationally efficient. Filter bank amplitude coefficients can then be obtained from the LP-derived spectrum as they were from the FT-derived spectrum [12].

A homomorphic system is useful for speech processing because it offers a methodology for separating the excitation signal from the vocal tract shape [12]. One space that offers this property is the cepstrum, computed as the inverse discrete Fourier transform of the log energy [12]. Cepstral coefficients are computed from the filter bank amplitudes through the following equation:

$$c(n) = \frac{1}{N_s} \sum_{k=0}^{N_s} \log|S_{avg}(k)| e^{j\frac{2\pi}{N_s}kn}, 0 \le n \le N_s - 1, \quad (3)$$

where $S_{avg}(k)$ is the average signal value in the k^{th} filter channel. In practice, the discrete cosine transform is used for computational efficiency. The cepstral coefficients are often weighted to minimize the non-information bearing variabilities, this process is known as liftering [8]. It is interesting to note that the speech recognition literature lists speaker characteristics as one such non-information bearing variability to suppress [8], but speaker recognition systems also use liftering [7].

The traditional linear predictive model approximates the power spectrum equally well at all frequencies in the analysis band. This is inconsistent with human hearing, which is nonlinear above 800 Hz. Perceptual linear predictive (PLP) analysis addresses this through critical band spectral resolution. PLP also addresses our ear's nonequal sensitivity at different frequencies with equal-loudness preemphasis [6]. Preemphasis can also used by cepstral front-ends by applying a simple high-pass filter to the signal before windowing [12]. Finally, PLP simulates the nonlinear relation between the intensity of a sound and its perceived loudness. These three aspects of PLP allow a small number of predictor coefficients to more closely model the way the human auditory system perceives the signal [6].

Both speech and speaker recognition systems bring short-term temporal information to the models by taking time derivatives of the base features [8, 7]. For example, a vowel sound can be recognized by finding its formants in the spectrum, but a consonant is modeled through the transition of the spectrum. The first order time derivative of the features are known as the delta coefficients, the second order derivative as acceleration or delta-delta coefficients. The time derivative is approximated through a regression formula drawing in a set number of frames before and after the current frame.

Speaker recognition systems also employ a feature selection module into the pattern recognition framework. For speech recognition, the entire signal must be mapped to a textual representation, but a speaker recognition system need not operate under this constraint. Hence the feature selection module saves only features corresponding to voiced sounds. Voiced sounds directly comply with the linear predictive modeling assumptions and are less effected by acoustic noise [1]. Further methods to reduce the effects of channel noise are discussed in the next section.

4. CHANNEL NORMALIZATION

The environmental differences between training and testing conditions cause a severe degradation in system performance for both speech and speaker recognition systems. Mammone et al. asserts that this "mismatched conditions" problem directly impedes the commercial deployment of the technology [7].

Variation of signal characteristics between trials is controlled through normalizing the input features. A constant yet unknown channel transform function affects the mean of the cepstral features. Since we are in the log domain this distortion can be considered an additive component to clean speech. This channel effect can be removed through cepstral mean subtraction (CMS), where the mean of each parameter is found over a number of frames and then subtracted from each frame [7]. Additive noise also results in a distribution with reduced variance. Variance normalization can be applied to cause each output feature to have unit variance regardless of the dynamic range of the input feature stream [5].

$$y_k(t) = \frac{x_k(t) - \overline{x_k}(t)}{\hat{\sigma}_k(t)} \tag{4}$$

Large vocabulary continuous speech recognition systems generally operate on short utterances of speech, ranging from five to twenty seconds. While some improvement in recognition accuracy is shown with utterance level cepstral mean and variance normalization, modeling the channel with an entire conversation's worth of data is more effective [13].

For both speech and speaker recognition cepstral mean subtraction improves system performance when the training and testing conditions are different. If the same channel is used for both training and testing, though, CMS will degrade speaker recognition accuracy. This is because CMS forces the long-term cepstral mean of the signal to be zero, throwing away some information about the speaker. More complicated strategies such as pole-filtered cepstral mean subtraction can achieve an LP model of the non-speech channel noise for increased robustness [7].

In addition to differing acoustic channel conditions, the same speakers voice can vary over time because of changes in health, emotional state, and age [11]. These ordinary human changes are beyond the scope of this paper, but Campbell's tutorial provides an extensive reference list on these topics [1].

5. EXPERIMENTS

A public domain feature extraction library, driven by **isip_transform**, is currently under development from the Institute for Signal and Information Processing. Pilot experiments have been run on a subset of the TIDIGITS corpus to exercise many of the concepts described in this paper. The training set consists of 500 utterances for a total of 860 seconds of audio data. The test set is comprised of 77 utterances. Due to the extremely small test set size the numbers reported must not be given much significance.

The small amount of data was chosen to facilitate rapid experiment turnaround, not a comprehensive evaluation. Additionally, TIDIGITS recognition is an isolated word recognition task with a very small vocabulary, so many acoustic problems can be overcome since the language model can tightly constrain the search space. Finally, the corpus was collected in a studio under controlled microphone conditions, not over the telephone network. The data set is useful, however, for quick algorithm evaluations.

The **isip_transform** utility was configured to extract many variations of the standard Mel-frequency cepstral coefficients (MFCCs) and filter bank amplitude coefficients (FBAs). An automated tutorial script is available to perform an entire recognition experiment, from feature extraction to hypothesis scoring. The script first runs the front-end program to extract features for both the training and test set. Next it uses the Baum-Welsh utilities to train word-level acoustic models based on our training features. Finally it decodes the test utterances and evaluates the hypothesis generated to produce a word error rate (WER) score. All experiments are described in Table 1.

Unless otherwise noted, a pre-emphasis filter $H(z) = 1 - 0.97z^{-1}$ was applied to the signal before finding the frequency response (energy is computed from the raw data). An normalized energy coefficient is included in each feature set. Cepstral mean subtraction was only applied when specified in the description. For all experiments a Hamming window function was used, and each window of data was normalized through mean subtraction across the single window.

FT-FBA coefficients are found by applying the Fourier Transform to a window of speech data and applying 24 overlapping triangular filter banks evenly spaced along the mel scale, recording the average spectral magnitude in each channel. FT-MFCC coefficients are found by taking the cepstrum of these FT-FBA coefficients. The LP-MFCC coefficients were obtained by first finding 12 predictor coefficients through the autocorrelation method, applying the Fourier Transform to the predictor coefficients to obtain a magnitude spectrum, then applying filter banks and the cepstrum to the LP-derived spectrum. "D" and "A" stand for delta and acceleration, respectively, and imply that differentials of the energy coefficient are also included.

It should be noted that the evaluation framework did not easily allow for modifying the size of the final feature vector. It was possible to remove the time derivative features, though, through duplicating coefficients. To remove both delta and acceleration coefficients, the base 13 coefficients were included in the final vector 3 times for a total of 39 elements. To remove only the acceleration coefficients, the

Exp #	Description	WER
exp_06	FT-MFCC + CMS + D + A no pre-emphasis	0.4%
exp_07	FT-MFCC + CMS + D + A,	0.8%
exp_100	FT-MFCC + D + A	0.8%
exp_101	FT-MFCC + D	0.8%
exp_108	FT-MFCC	7.9%
exp_103	LP-MFCC + D + A	0.8%
exp_106	LP-MFCC + D	2.4%
exp_102	LP-MFCC	6.7%
exp_104	FT-FBA	100%
exp_107	FT-FBA + D	100%

Table 1: TIDIGIT results

base 13 coefficients were duplicated once to overwrite the removed coefficients. This may add additional weight to the base coefficients, but the powerful statistical methodology provided through the use of hidden markov models should compensate for such variability.

As suggested by Table 1, linear predictive cepstral coefficients do behave similarly to Fourier Transform MFCCs. This is to be expected, since the main drawback of the predictive approach is breakdown in noisy conditions. Removing all temporal coefficients causes performance to degrade by an order of magnitude, the first order delta coefficients having a stronger effect than the acceleration coefficients. It is interesting to note that for FT-derived coefficients the acceleration coefficients appear to be useless, but for LP-coefficients they do have an effect. This is most likely a statistical anomaly.

Also, the cepstral mean subtraction normalization technique had no effect on the overall system performance. This is probably because the channel conditions do not really need to be normalized for this corpus. It is interesting to note that pre-emphasis actually degrades performance on this data, this is most likely a statistical anomaly.

The filter bank amplitude experimental results are disappointing. Feature files were properly generated, but the scale of each coefficient was very different than the cepstral coefficients. It is unlikely that a range problem would cause 100% error, though, so most likely a fatal error occurred somewhere in the evaluation framework, as little of the speech recognition toolkit has been verified to work with anything except the standard 39 dimensional MFCC vector. Further study is necessary to track down the problem, possibly the addition of mean and variance normalization to the filter bank amplitude coefficients could result in a working system.

6. SUMMARY

Both speech and speaker recognition systems perform pattern recognition on speech data using similar feature extraction algorithms. These algorithms are based on acoustic models of the vocal tract and ear canal.

The most prevalent features used in speech and speaker recognition are mel-frequency cepstral coefficients. These coefficients are extracted from the signal nearly as fast as human articulators can move from one position to another, 100 times a second. Cepstral coefficients can be obtained through the Fourier Transform or linear predictive analysis. Perceptual linear predictive analysis is an extension of LP modeling which brings more physiological knowledge of the human auditory and articulatory systems to bear on the problem. One significant difference in speech and speaker recognition systems is that speaker recognition systems can employ a feature selection module to select only voiced sounds.

For a small subset of the TIDIGIT corpus both the linear predictive and Fourier Transform derived cepstral coefficients performed very well. The system behaved identically with full temporal coefficients. This is due to the clean channel conditions of the corpus.

The accuracy of both speech and speaker recognition systems degrade significantly when the environmental conditions vary between system training and testing. Cepstral mean subtraction is a normalization technique used in both application domains. Creating voice processing systems robust to noise and other environmental variability would greatly enhance the technology's commercial appeal.

REFERENCES

[1] Campbell, J.P., Jr. 1997. Speaker recognition: A tutorial. In *Proceedings of the IEEE* 85 (9): 1437-62.

[2] Cole, Ron, and Victor Zue, ed. 1996. Spoken language input. In *Survey of the state of the art in human language technology*, chapter 1. http://cslu.cse.ogi.edu/HLTsurvey/HLTsurvey.html (Accessed 17 February 2000).

[3] Deshmukh, Neeraj, Aravind Ganapathiraju, and Joseph Picone. 1999. Hierarchical search for large

vocabulary conversational speech recognition. *IEEE Signal Processing Magazine* 16 (5): 84-107.

[4] Frischholz, Robert W., and Ulrich Dieckmann. 2000. BioID: A multimodal biometric identification system. *Computer* 33 (2): 64-8.

[5] Haeb-Umbach, Reinhold, Xavier Aubert, Peter Beyerlein, Dietrich Klakow, Meinhard Ullrich, Andreas Wendemuth, and Patricia Wilcox. 1998. Acoustic modeling in the Philips Hub-4 continuous-speech recognition system. In *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop held at the Lansdowne Conference resort, Lansdowne, Virginia on February 1998*, by the National Institute of Standards and Technology.

[6] Hermansky, Hynek. 1990. Perceptual linear predictive (PLP) analysis of speech. In *Journal of the Acoustical Society of America* 87 (4): 1738-52.

[7] Mammone, R. J., Xiaoyu Zhang, and R. P. Ramachandran. 1996. Robust speaker recognition: A feature-based approach. *IEEE Signal Processing Magazine* 13 (5): 58-71.

[8] Mantha, Vishwanath, Richard Duncan, Yufeng Wu, Jie Zhao, Aravind Ganapathiraju, Joseph Picone. 1999. Implementation and analysis of speech recognition front-ends. In *Proceedings of the IEEE Southeastcon held in Lexington, Kentucky on March 1999*. The institute for Electrical and Electronic Engineers. 32-5.

[9] Markowitz, Judith. 1998. Is that really you? *Byte* 23 (6): 123.

[10] Pankanti, Sharath, Ruud M. Bolle, and Anil Jain. 2000. Biometrics: The future of identification. *Computer* 33 (2): 46-9.

[11] Phillips, P. Jonathon, Alvin Martin, C.L. Wilson, Mark Przybocki. An introduction to evaluating biometric systems. *Computer* 33 (2): 56-63.

[12] Picone, Joseph. 1993. Signal modeling techniques in speech recognition. In *Proceedings of the IEEE* 81 (9): 1215-47.

[13] Sundaram, Ramasubramanian, Aravind Ganapathiraju, Jonathan Hamaker, and Joseph Picone. 2000. The ISIP 2000 Hub 5E conversational speech evaluation system. To be presented at the *Speech Transcription Workshop at the University of Maryland University College in College Park, Maryland on May 2000*, by the National Institute for Standards and Technology.

[14] Wu, Chung-Hsien, and Jau-Hung Chen. 1997. Speech activated telephony email reader (SATER) based on speaker verification and text-to-speech conversion. *IEEE Transactions on Consumer Electronics* 43 (3): 707-716.

[15] Young, Steve. 1996. A review of large-vocabulary continuous-speech recognition. *IEEE Signal Processing Magazine* 13 (5): 45-57.

[16] Zhao, Jie. 2000. Overlap in the cepstral space. In *Fundamentals of Speech Recognition*, lecture 2. Mississippi State University. http://www.isip.msstate.edu/publications/courses/ece_8463/lectures/current/ (Accessed 15 April 2000).