Speaker Identification Using Gaussian Mixture Models

Pawan Kumar
Electronics & Communication Engg. Department, B.I.T., Mesra, Ranchi, India

Mahesh Chandra
Electronics & Communication Engg. Department, B.I.T., Mesra, Ranchi, India

Abstract—In this paper, the performance of Perceptual Linear Prediction (PLP) features has been compared with the performance of Linear Prediction Coefficient (LPC) features for speaker identification. Two classification techniques, Gaussian Mixture Models (GMM) and Vector Quantization (VQ) with Dynamic time wrapping (DTW) are used for classification of speakers based on their speech samples into respective classes. A database of fifty speakers, twenty one males and twenty nine females has been prepared in clean environment. The identification performance of PLP features is 3.6% better than LPC features with VQ and DTW classifier. PLP features have also shown 1.2% increment in identification performance over LPC features with GMM classifier.

Keywords—Gaussian Mixture Models, Perceptual Linear Prediction, Linear Prediction Coefficient, Vector Quantization

I. INTRODUCTION

SPEECH processing is a diverse field with many applications but here speaker recognition will be discussed. Speaker recognition can be divided into two classes, text dependent and text independent speaker recognition [1][2][3][4]. Text-dependent speaker recognition systems require the speaker to produce speech for the same text in both training and testing, whereas in text-independent speaker recognition, the text during training and testing are different. Text-independent speaker recognition systems are based on factors such as the shape and size of the vocal tract, dynamics of the articulators, rate of vibration of the vocal folds, accent imposed by the speaker and speaking rate. All these factors are reflected in the speech signal and hence are useful for speaker recognition. The variation in the size and shape of the vocal tract from one speaker to another is reflected as the differences in the resonance frequencies of the short-time spectrum envelope of the speech signal. Speaker verification and identification [2][3] are two broad classes of speaker recognition which are used in many applications. Speaker identification is the process of determining to which of the registered speakers a given utterance belongs. Speaker verification is the process of accepting or rejecting the identity claim of the speaker.

In this paper Gaussian mixture model [1][4] and VQ with DTW [5] classifiers are used for classifying speakers into their respective classes. Prior to construction of GMM for each speaker, speech signal is first transformed into a set of spectral vectors which is a convenient representation of a person’s vocal tract structure and would constitute an important factor distinguishing one person’s voice from another. LP based features [2][6][7][8], PLP [9] feature sand MFCC[8] features have been used by various researchers for speech recognition, speaker recognition and languages identification purposes. Feature extraction techniques are given in Section 2. Experimental setup and results are explained in Section 3. Finally conclusions are drawn in Section 4.

II. FEATURE EXTRACTION TECHNIQUES

The raw speech signal is complex and may not be suitable for feeding as input to the automatic speaker identification system; hence the need for a good front-end arises. The task of this front-end is to extract all relevant acoustic information in a compact form compatible with the acoustic models. In other words, the preprocessing should remove all non-relevant information such as background noise and characteristics of the recording devices, and encode the remaining (relevant) information in a compact set of features that can be given as input to the classifier. Features can be defined as a minimal unit, which distinguishes maximally close classes. Preprocessing stage is common in both the feature extraction techniques LPC [2][5] and PLP [9] as shown in Fig. 1. Pre-emphasis filtering, normalization and mean subtraction are the three steps in pre-processing. By applying a pre-emphasis filter the glottal waveform and lip radiation characteristics are eliminated. Due to possible mismatch between training and test conditions, it is considered good practice to reduce the amount of variation in the data that does not carry important speech information as much as possible. For instance, differences in loudness between recordings are irrelevant for recognition. For reduction of such irrelevant sources of variation, normalization transforms
are applied. During normalization every sample value of the speech signal is divided by the highest amplitude sample value. Mean of the utterance is subtracted from the utterance to remove the DC offset and some of the disturbances induced by recording instruments. Framing is done to make the speech signal stationary. The pre-emphasized speech signal is blocked into frames of 240 samples; with adjacent frames being separated by 144 samples.

A. Linear Prediction Coefficient

The human speech production process [3] reveals that the generation of each phoneme is characterized basically by two factors: the source excitation and the vocal tract shaping. In order to model speech production we have to model these two factors. To understand the source characteristics, it is assumed that the source and the vocal tract model are independent. The vocal tract model $h(n)$ is excited by a discrete time glottal excitation signal $u(n)$ to produce the speech signal $s(n)$ shown in Equation (1).

$$s(n)=h(n)\ast u(n) \quad (1)$$

Linear prediction technique [5] [6] [7] is used to derive the filter coefficients (corresponding to the vocal tract) by minimizing the mean square error between the input and the estimated sample. These coefficients are extracted using the auto-correlation method or the covariance method. The all pole representation of the vocal tract transfer function $H(z)$ can be represented by Equation (2).

$$H(z) = \frac{G}{A(z)} = \frac{G}{1+2\alpha z^{-1}+2\alpha z^{-2}+...+\alpha^p z^{-p}} \quad (2)$$

The next step in the processing is to windows each individual frame so as to minimize the signal discontinuities at the beginning and end of each frame. To minimize this, Hamming window is used. In frequency domain spectrum of the hamming window is more smoother than any other window.

B. Perceptual Linear Prediction (PLP)

PLP provides a representation corresponding to a smoothed short-term spectrum that has been compressed and equalized much as done in human hearing. It can be assumed similar to mel-cepstrum based features. In PLP technique, several well-known properties of hearing are simulated by practical engineering approximations, and the resulting auditory like spectrum of speech is approximated by an autoregressive all-pole model. PLP provides reduced resolution at high frequencies that indicates auditory filter bank based methods, yet provides the orthogonal outputs that typify cepstral analysis. PLP uses linear predictions for spectral smoothing; hence the name is perceptual linear prediction. The different steps of PLP analysis are as follows.

I. Power spectral estimate for the windowed speech signal is computed. This is done by windowing the analysis region with Hamming window, calculating the FFT and computing its squared magnitude.

II. The power spectrum within overlapping critical band filter responses is integrated. For PLP, trapezoidal shaped filters are applied at 1-bark intervals, where the bark axis [9] is derived from the frequency axis by using a warping function from Schroeder, given in Equation 3. The Bark scale is linear at low frequencies and logarithmic at high frequencies.

$$\zeta(\omega) = 6 \ln \left( \frac{\omega}{1200\pi} + \left[ \frac{\omega}{1200\pi} \right]^2 + 1 \right) \quad (3)$$

Where $\omega$ is angular frequency in radians/second. This effectively compresses the higher frequencies into a narrow band. The critical band masking, symmetric frequency domain convolution on the Bark warped-
frequency scale then allows low frequencies to mask the high frequencies while at the same time smoothing the spectrum an effect consistent with the psycho-acoustic results.

III. The spectrum is pre-emphasized to approximate the unequal sensitivity of human hearing at different frequencies. This step is implemented as an explicit weighting of the elements of critical band spectrum.

IV. Spectral amplitude is compressed. The effect of this step is to reduce amplitude variations for the spectral resonances.

V. An inverse DFT is performed. As a result of this step autocorrelation coefficients are obtained, but these coefficients are from a compressed spectrum. Since the power spectrum values are real and even, only the cosine components of inverse DFT are to be calculated.

VI. Spectral smoothing is performed. This step is done by solving the autoregressive equations constructed from the autocorrelations of the previous step.

VII. The autoregressive coefficients are converted to cepstral variables.

III. EXPERIMENTAL RESULTS AND DISCUSSION

A database of fifty speakers, twenty one males and twenty nine females, for a total of ten isolated Hindi digits (‘shunya’, ‘ek’, ‘do’, ‘teen’, ‘chaar’, ‘paanch’, ‘chheh’ ‘saat’, ‘aath’ and ‘nao’) has been prepared with sampling frequency of 16 kHz and 16 bits per sample. Ten samples of each Hindi digit for all fifty speakers were recorded in order to prepare the database. Speakers from different social classes and of different age groups (18-26 years) were chosen. The experimental set-up of speaker identification system using VQ with DTW classifier is shown in Fig. 2.

By using LPC and PLP techniques, features were extracted for all 10 utterances of all digits for all 50 speakers. In order to have more temporal information, the duration of each utterance was divided into number of sub-frames. The sub frames were of 15ms with 6ms overlapping. For both kinds of feature extraction techniques sixteen features are calculated for each frame. All feature vectors of all frames of an utterance were coded into single feature vector using VQ. In this way 5000 feature vectors were extracted for 50 speakers of all Hindi digits and these were stored for further use during training. Further the hundred feature vectors of each speaker were coded into a single feature vector using VQ. Finally a total of 50 feature vectors were received, one feature vector corresponding to one speaker. These were stored as 50 reference templates corresponding to 50 speakers to be used during testing. Feature vector of second sample of each digit of each speaker was taken for testing.

During testing phase, total 500 feature vectors, ten for each speaker were tested.

The experimental set-up of speaker identification system using GMM classifier is shown in Fig. 3.

By using LPC and PLP techniques, frame based features were extracted for all ten samples of all ten digits for all fifty speakers. Features vectors of respective speaker were used to prepare GMM for that speaker. In this way fifty GMMs were prepared for 50 speakers. During testing, feature vectors of first twenty frames of sixth sample of
each digit of each speaker were taken. In this way, total 10,000 features vectors, 200 for each speaker were tested. The identification results are shown in Fig. 4.

The identification performance of PLP features is 3.6% better than LPC features with VQ and DTW classifier. PLP features have also shown 1.2% increment in identification performance over LPC features with GMM classifier.

IV. CONCLUSION

As the LPC model is an all pole model, it can capture the resonant frequencies, or formants, but not the zeros, which are important for nasalized sounds. Due to this reason LPC features performs poor for nasalized sounds. One more reason for better performance of PLP features is that, these are cepstral features. Cepstral features are useful because they operate in a domain in which the excitation function and the vocal tract filter function are separable.

REFERENCES


