International Journal of Advanced Computer Research (ISSN (print): 2249-7277 ISSN (online): 2277-7970)
Volume-3 Number-3 Issue-11 September-2013

Speaker Verification Using I-Vectors

N.S.Kalkar[1], S.P.Savarkar[2], RajaniP.K[3]

Abstract

This paper deals with the study of low-variance multi-taper Mel-Frequency Cepstral Coefficient (MFCC) and Perceptual Linear Prediction (PLP) features in i-vector speaker verification. Hamming windowed periodogram spectrum estimate is a important method to calculate the MFCC and PLP features. Single tapered spectrum estimate has large variance, which can be reduced by averaging spectral estimates obtained using a set of different tapers, leading to a multitaper spectral estimate. The multi-taper spectrum estimation method has proven to be powerful especially when the spectrum of interest has a large dynamic range or varies rapidly. In this study primary goal is to validate those findings using an up-to-date i-vector classifier. Robust Perceptual Linear Prediction (PLP) features using multitapers. Sine – Weighted Cepstrom Estimator based multitaper method provides average relative reductions of 12.3% and 7.5% in Equal Error Rate, respectively. For the Multi-Peak Multi-Taper method, the corresponding reductions are 12.6% and 11.6%, respectively. Finally, the Thomson multitaper method provides error reductions of 9.5% and 5.0% in EER for MFCC and PLP features, respectively. Both the MFCC and PLP features computed via multitapers provide systematic improvements in recognition accuracy.

Keywords

Speaker verification, Multi-taper spectrum, Feature extraction, i-vectors, MFCC, PLP.

1. Introduction

From time immemorial, man has strived to flash "news and views" to fellow beings. Primitive man interacted through inarticulate utterances, mimics and gestures. Then evolved, quite naturally, the faculties of speech and languages, God's priceless boon to mankind, facilitating the expression of thought and emotions. The inherent superiority of speech over other modes of communication and its seamless integration with other services, makes it the favorite choice for providing easy –to-use, efficient and affordable linkage medium between man and machine.

The unique features offered by digital technology have made it very attractive for speech signal processing. The fusion of digital methods with human voice, one of the most complicated analog signals, has posed great challenges. Communication is the focal point of all activities ,whether it is person to person, machine to person or person to machine communication. The communication between human beings embraces the subject of coding and decoding of speech for its storage and efficient transmission. The communication from machine to person gives the machine a "mouth" with which it can deliver information to humans. In the same way, the communication from person to machine gives the machine an"ear" through which it can listen to instruction from humans, and carry out the necessary action.

SPEAKER VERIFICATION is the process of deciding whether or not an unknown speech specimen was spoken by the individual speaker whose identity was claimed. THE OBJECTIVE OF SPEAKER VERIFICATION is to unambiguously identify the talker attempting to get voice access to machine,e.g., to distinguish between genuine and false speakers.

A. Speaker Verification

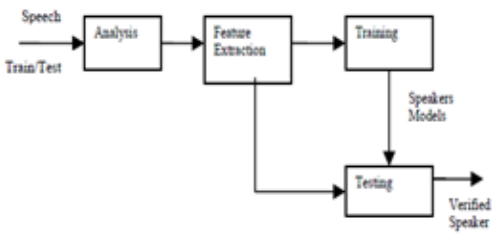


Figure 1: Speaker Verification System

Speaker verification is the process of accepting or rejecting the identity claim of a speaker. In most of the applications voice is used to confirm the identity claim of a speaker. Speaker recognition system may be viewed as working in four stages namely Analysis, Features Extraction, Modeling and Testing as shown in Figure.1.

Stages in the Development of Speaker Verification System. Speech data contains different types of information that conveys speaker identity. These include speaker specific information due to the vocal

tract, excitation source and behavioral traits. The speech signal is produced from the vocal tract system. The physical structure and dimension of vocal tract as well as the excitation source are unique for each speaker. This uniqueness is embedded into the speech signal during speech production and can be used for speaker verification. To obtain the good representation of these speaker characteristics, speech data needs to be analyzed. The speech analysis stage deals with the selection of suitable frame size and frame shift for segmenting the speech signal for further analysis and feature extraction.

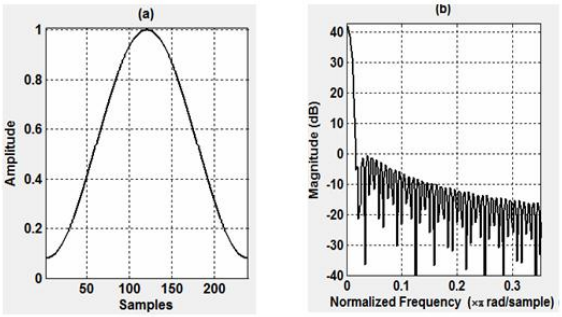
B. Present Method

Useful information extraction from speech has been a subject of active research for many decades. Feature extraction (or front-end) is the first step in an automatic speaker or speech recognition system. It transforms the raw acoustic signal into a compact representation. Since feature extraction is the first step in the chain, the quality of the subsequent steps (modeling and classification) strongly depends on it. The Mel-Frequency Cepstral Coefficient (MFCC) and Perceptual Linear Prediction (PLP) front-ends have been dominantly used in speech and speaker recognition systems and they demonstrate good performance in both applications. The MFCC and PLP parameterization techniques aim at computing the speech parameters similar to the way how a human hears and perceives sounds. Since these features are computed from an estimated spectrum, it is crucial that this estimate is accurate. Usually, the spectrum is estimated using a windowed periodogram via the Discrete Fourier Transformation (DFT) algorithm. Despite having low bias, a consequence of the data tapering (windowing) is increased estimator variance.

A windowed direct spectrum estimator is the most often used power spectrum estimation method in speech processing applications. For the m^{th} frame and k^{th} frequency bin an estimate of the windowed periodogram can be expressed as [6]:

$$\hat{S_d}(m, k) = \left| \sum_{j=0}^{N-1} W(j) s(m.j) e^{2\pi i k/N} \right|^2 , \quad \text{---- (1)}$$

Where k Î{0,1,...,K -1} denotes the frequency bin index,N is the frame length, s (m, j) is the time domain speech signal and w (j) denotes the time domain window function, also known as taper. The taper, such as the Hamming window, is usually symmetric and decreases towards the frame boundaries. Eq. (1) is sometimes called single-taper, modified or windowed periodogram. If w (j) is a rectangular or uniform taper, Eq. (1) is called a periodogram. Figure 2 presents time- and frequency-domain plot of the Hamming window [5].



Time Domain Frequency Domain

Figure 2: Hamming window, in (a) time domain, (b) frequency domain

C. Proposed Method

MFCC or PLP features computed from windowed periodogram estimated spectrum have also high variance. One elegant technique for reducing the spectral variance is to replace a windowed periodogram estimate with a multi-taper spectrum estimate [1].

A good estimater is one that makes some suitable tradeoff between low bias and low variance. Propose to use multi-taper MFCC and PLP features in an i – vector speaker verification system. I – Vector do a good job in compensating for variability's in the speaker model space.

Multitaper methods outperform the conventional periodogram technique. The multi-taper spectrum estimation method has proven to be powerful especially when the spectrum of interest has a large dynamic range or varies rapidly.

The idea behind multi-tapering is to reduce the variance of the spectral estimates by averaging M direct spectral estimates, each with a different data taper. If all M tapers are pairwise orthogonal and properly designed to prevent leakage, the resulting multi-taper estimates outperform the windowed periodogram in terms of reduced variance, specifically, when the spectrum of interest has high dynamic range or rapid variations. Therefore, the variance of the MFCC and PLP features computed via this multitaper spectral estimate will be low as well.

In the multi-taper method, only the first of the data tapering windows has the traditional shape. The spectra from the different tapers do not produce a common central peak for a harmonic component. Only the first taper produces a central peak at the

harmonic frequency of the component. The other tapers produce spectral peaks that are shifted slightly up and down in frequency. Each of the spectra contributes to an overall spectral envelope for each component [1].

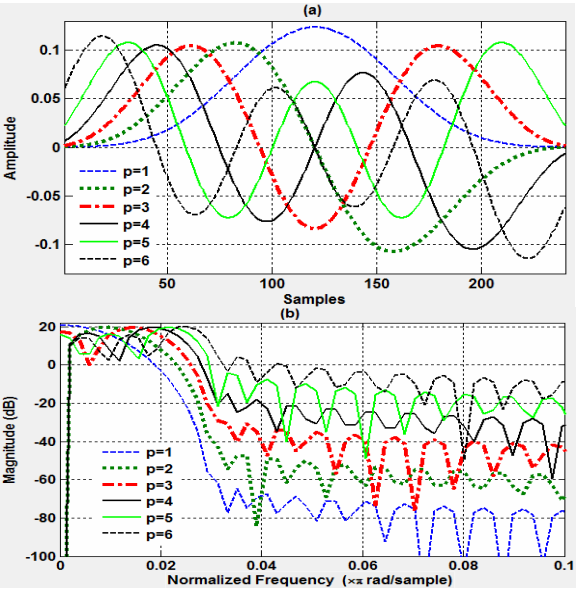


Figure 3: Thomson multi-tapers for M = 6 in (a) time and (b) frequency domains

The objective of the taper is to prevent energy at distant frequencies. Various multi-taper methods have been proposed for spectrum estimation, such as Thomson multitaper, Sinusoidal Weighted CepstrumEstimator multi-taper and Multipeak multi-taper. In this seminar Thomson multi-tapers method is used.

Speaker verification accuracy when the standard windowed periodogram was replaced by the Thomson multi-taper. Instead of simply averaging (using uniform weights) the individual spectral estimates in forming the multitaper estimate, weighted averaging (using non-uniform weights) improves performance.

Only the first taper (p = 1) in the multi-taper method produces a central peak at the harmonic frequency of the component while the other tapers (p > 1) produce spectral peaks that are shifted slightly up or down in frequency. The information lost at the extremes of the first taper is included and indeed emphasized in the subsequent tapers. As can be seen from Fig. 4, attenuation in the side-lobes decreases with each taper in the sequence, i.e., spectral leakage increases for the higher-order tapers. If uniform weights are applied to get the final spectrum estimate, the energy loss at higher-order tapers will be high. In order to compensate for this increased energy loss, a weighted average (using non-uniform weights) is used instead of simply averaging the individual estimates. In the weights are changed adaptively to optimize the bias/variance trade off of the estimator.

2. Speech Processing

A. Production Of Speech

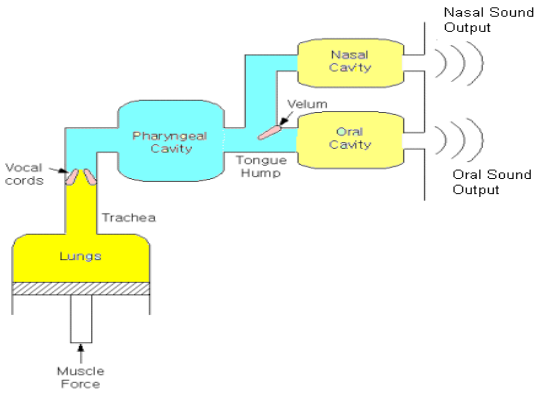


Figure 4: Human Voice Production System

Vocal cords and vocal tract are together responsible for the human speech. The vocal cords, consisting of two bands of tough, elastic tissue, located at the opening of the larynx, vibrate when the air from the lungs passes between them producing sound waves which are emitted from the lips and to some extent from the nose; these sound waves are heard as speech. The vocal tract includes the larynx, the pharynx and the nasal cavity. The larynx is the upper part of the trachea containing the vocal cords, and pharynx is the passage way from the mouth to the throat.

B. Voiced and Unvoiced Sound

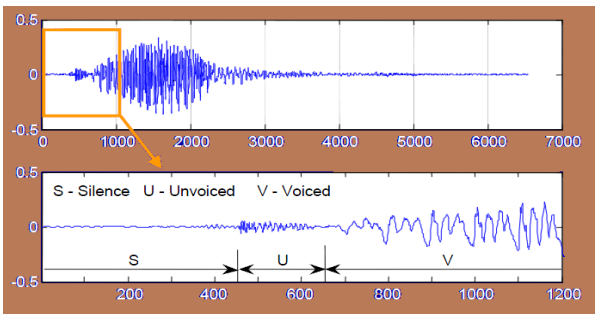


Figure 5: Voiced and Unvoiced Sound Signal

Speech sounds are of two kinds: Voiced Sounds and Unvoiced or Fricative Sounds. Voiced Sounds are produced by quasi-periodic pulses of air exciting the vocal tract. The pulses, in turn, are generated by forcing air through the epiglottis, a small flap of tissue situated at the entrance to the larynx, with tension of the vocal cords adjusted in such a way they are thrown to the larynx, with the tension of the vocal cords adjusted in such a way that they are thrown into a relaxation oscillation. To produce an unvoiced sound, a constriction is formed at some point along the vocal tract, generally towards the mouth. Air is forced through the constriction at a high speed resulting in turbulent air flow. A broad-spectrum noise source is created which excites the vocal tract leading to the emission of an unvoiced sound.

C. Digitization Of Speech

The acoustic energy of speech is converted into electrical energy by suitable elecroacoustic transducer like the moving –coil, electrostatic or piezoelectric microphone. In preparation for performing signal processing operations on the speech signal, the speech waveform is passed through a low pass filter restraining the components below a frequency W Hz.It is then sampled at a rate >2W, the Nyquist rate, quantized and covered into binary digits. Although the frequency W is determined by the intended application, it usually lies between 3 and 5KHz. Straightforward application of sampling theorm shows that a sampling rate >8000 samples per second is adequate for digital representation of telephone bandwidth ~ 4KHz speech signals. An encoding rate of 16 bits per second is followed.
The digitized signal is ready for processing and various operations can be carried out on the digitized signal, as demanded by the problem at hand.

D. Speech Coding

Coding of speech embraces the different processes by which, the speech signal is represented in digital form, preferably at a low bit rate, within the acceptable intelligibility and quality limits prescribed by the application in question. By "intelligibility of speech" mean the ease with which it can be understood by a listener. Quality of speech refers to how natural it sounds as compared to the words uttered by a human being.

Intelligibility and quality of speech are measured in terms of a performance index called Mean Opinion Score. To determine the MOS of a speech source, a large number of listeners are requested to rate the given samples of speech and allot marks as follows: Excellent -5, Good-4, Fair-3, Poor-2 and Bad-1 These values, after averaging, yields the MOS score for the coder. For a high –quality coder, the MOS values range between 4.0 and 4.5. Thus the MOS procedure assigns numerical values to subjective evaluation. Following table discuss the principal parameters of a speech coder indicating the properties measured by them. Waveform coding ,Baseband coding and narrow band coding are deserve special methods of speech coding. The "Ideal Speech Coder" gives a high quality of speech at a low bit rate. It also produces a small delay and has a low complexity level. Practical coders seek a trade-off among these four attributes.

Table 1: Parameter of Speech Coder[V.K. Khanna, "Digital Signal Processing"]

Sr.No	Parameter	Property Measured
1	Bit Rate	Measures the number of special properties of speech exploited.
2	Quality	A measure of the amount of speech signal , expressed in terms of Intelligibility and naturalness of speech.
3	Delay	A measure of amount of speech signal necessary in order to determine the parameters of a speech coder reliably. IT equals the sum of delays experienced by the signal in transit through the coder and the decoder i.e., it is equal to Encoder Dealy + Decoder Dealy.
4	Complexity	Measures the computational requirement for coder implementation in DSP Hardware.
5	Pitch	Measures the frequency of sound in Hz
6	Loudness	Measures the Loudness of sound in dB

E. Feature Extraction

Feature extraction is the process that extracts a small amount of data from the voice signal that can later be used to represent each speaker. It should have high computation efficiency. Feature extraction should Scale and Shift Invariant as well as Noise Robust.The speech feature extraction in a categorization problem is about reducing the dimensionality of the input-vector while maintaining the discriminating power of the signal. Speaker verification systems, that the number of training and test vector needed for the classification problem grows exponential with the dimension of the given input vector, so we need

feature extraction. But extracted feature should meet some criteria while dealing with the speech signal.
Such as:

- Easy to measure extracted Speech features.
- Distinguish between speakers while being intra speaker variability's.
- It should not be susceptible to mimicry.
- It should show little fluctuation from one speaking environment to another.
- It should be stable over time.
- It should occur frequently and naturally in speech.

MFCC

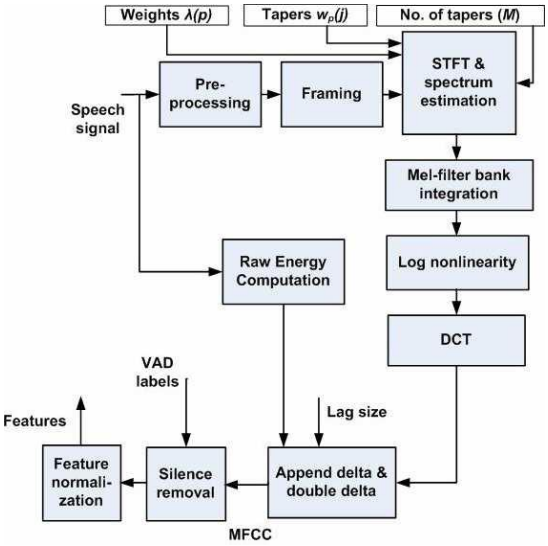


Figure 6: Mel Frequency Cepstral Coefficients (MFCC)

As shown in Figure 6 speech signal is input signal giving to the pre-processing block and raw energy computation block. In Pre-processing block DC is removed and pre-emphasis is done using 1st order high-pass filter with transfer function $H(Z) = 1 - 0.97 * Z^{-1}$.
Second step is Framing & Windowing: In this step the speech signal is blocked in to frames of N samples, with adjecent frames being seperated by M (M < N). Windowing is done with individual frame so as to minimize the signal discontinuties at the beginning and end of each frame, typically the Hamming window is used [1].
Third step is STFT ie Short Time Fourier Transform, this analysis is carried out using a single taper or multitaper technique.Forth step is Mel Filter Bank Integration, this is performed for ouditory spectral analysis. Passes filter outputs through log function then logarithmic non linearity stage follow.Fifth step is DCT i.e .Discrete Cosine Transform, DCT is technique for converting signal into elementry comoponent. DCT is similar to DFT since it decomposes a signal into series of harmonic cosine fuction.
Sixth steps is Append delta and double delta, these are the 1st and 2nd order time derevatives.
Seventh steps is silence removal, In this step silence at the beginning and end of the speech samples will be removed.
Last block is feature normalization, in this step short time gausianisation is used. They had a property to minimize the rise and fall time of step function.
It has a minimum group delay. Gaussian filter modifies the input signal by convolution with a gaussian function.

Features of MFCC:

- Temporal variation or time derivative features
- Spectral Transition plays an important role in human speech production.
- Not sensitive to slow channel-dependent variations of static parameters
- First order difference is affected by various types of noise, thus smoothing necessary polynomial expansion of time derivatives.
- Second order derivatives: acceleration also often used

Typical set of parameters: MFCC, ΔMFCC and ΔΔ MFCC.

PLP

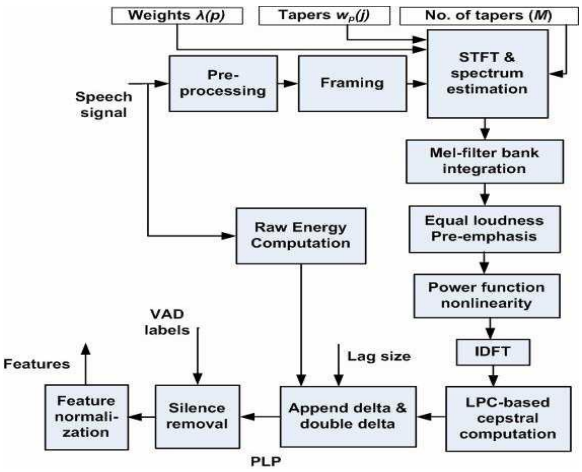


Figure 7: Perceptual linear prediction (PLP)

As shown in Figure 7 speech signal is input signal giving to the pre-processing block and raw energy computation block. In Pre-processing block DC is removed and pre-emphasis is done using 1st order

high-pass filter with transfer function H(Z) = 1- 0.97* Z^{-1} [1].

Second step is Framing & Windowing: In this step the speech signal is blocked in to frames of N samples, with adjecent frames being seperated by M (M < N). Windowing is done with individual frame so as to minimize the signal discontinuties at the beginning and end of each frame, typically the Hamming window is used.

Third step is STFT i.e. Short Time Fourier Transform, this analysis is carried out using a single taper or multitaper technique.Forth step is Male Filter Bank Integration, this is performed for ouditory spectral analysis. Passes filter outputs through log function then logarithmic non linearity stage follow.

Fifth step is Equal Loudness pre-emphasis. Pre-emphasis is performed based on an equal loudness curve after frequency integration.Sixth step is Power Function Nonlinearity-Based on power law.

Seventh step is IDFT.In this step obtaining a perceptual auto correlation sequence ,following the linear prediction analysis.Eight step is LPC based cepstral computation.This is done to obtain the final features from Linear Prediction.

Features of PLP:

- Critical band spectral resolution.
- Equal loudness curve
- Intensity loudness power low
- Computationally efficient and yields a low dimentional represention

F. I-Vector Framework

Extraction of I – Vector

I-vector extractors have become the state-of-the-art technique in the speaker verification field. An i-vector extractor represents entire speech segments as low-dimensional feature vectors called i-vectors. In this paper, however use a gender-independent i-vector extractor, as shown in Figure. 8, trained on both microphone and telephone speech. The universal background model (UBM) used in this i-vector extractor is also gender-independent. The advantage of a gender-independent system is simplified system design as separate female and male detectors do not need to be constructed. In order to handle telephone as well as microphone speech, the dimension of the i-vectors is reduced from 800 to 200 using ordinary Linear Discriminant Analysis (LDA). The purpose of applying length normalization is to Gaussianize the distribution of the i-vectors so that a simple Gaussian PLDA model can be used instead of the heavy-tailed PLDA model. A heavy-tailed PLDA is 2 to 3 times slower than the Gaussian PLDA [1].

Gender-independent i-vector extractor is of dimension 800. After training the gender-independent UBM, train the i-vector extractor using the Baum-Welch (BW) statistics. Baum-Welch algorithm is used to find the unknown parameter of a Hidden Markov Model (HMM).

It is a expectation – Maximization algorithm. It can compute maximum likelihood estimates.

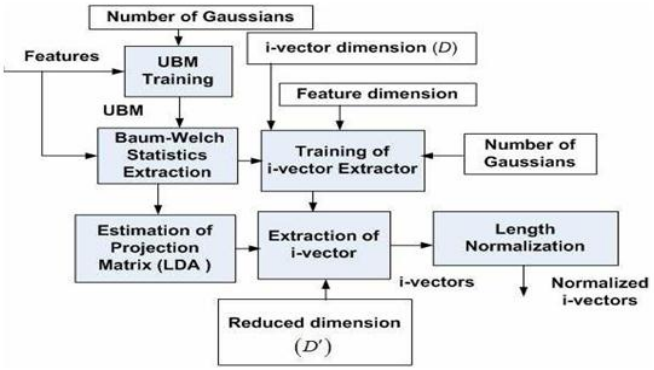


Figure 8: Block diagram of Gender-independent i-vector extractor

Features of i - Vector:

- State of the art technique in the speaker verification.
- Convert an entire speach recording into low dimentional feature.
- i - vectors are the identity vectors
- i- vector extractor are suitable for speaker verification with both microphone and telephone speech.
- i-vector are smaller in size to reduce the execution time of the recognition task, while maintaning recongnition performance

G. Generative PLDA Model

Construct a speaker detector using a generative PLDA model for a pair of i-vectors in a trial. The model assumes that the i-vectors were produced by simple random processes. In the model, the pair of i-vectors Z1 , Z2 is produced as follows:

$$Z1 = Vy1 + X1$$
$$Z2 = Vy2 + X2 \text{---------} (2)$$

Where the hidden speaker variables, y1 , y2 , are d-dimensional vectors sampled from a continuous multivariate between speaker distribution. y1 = y2 , for target trials whereas for nontarget trials, y1 and y2 are sampled independently. The hidden channels, x1 , x2 are D-dimensional and sampled from a continuous multivariate within-speaker distribution.

Normally d ≤ D, but here d = D. The between- and within-speaker distributions are either normal or heavy-tailed, the d ×D matrix V is a fixed hyper-parameter and Z1 , Z2 are observed variables. There are two types of hidden variables: (i) the continuous nuisance variables: x1 , x2 , y1 , y2 and (ii) the variable of interest to be inferred, i.e., the trial type, which can have the discrete values target (T) or non-target (N) [2].

Gender modeling

Let, g1 , g2 represent the genders of the speakers that produced Z1 , Z2 which take the values male (M) or female (F). For a target trial, g1 = g2 ,while for a nontarget trial they ma y be different.

The generative model needs priors for all hidden variables. The priors for the continuous hidden variables are within and between speaker distributions mentioned above. In this paper the prior for the trial type is not needed. The priors for the gender labels are trial-type dependent and are defined as [2]:

$P_M = P(MM \mid T)$ $P_F = P(FF \mid T)$ ---------- (3)

$Q_{MM} = P(MM \mid N)$ $Q_{FF} = P(FF \mid N)$ --------- (4)

$Q_{MF} = P(MF \mid N)$ $Q_{FM} = P(FM \mid N)$, ------ (5)

Where the event g_1 = M and g_2 = F is denoted by MF, $P_M + P_F = 1$, and $Q_{MM} + Q_{FF} + Q_{FM} = 1$. Equiprobable priors are used for this case. These priors take values of 0 or 1 in the limiting case of given gender labels.

Scoring:

For gender-independent scoring of the model, assume to have the following available gender-dependent likelihoods:

For targets:

$P(Z_1, Z_2 \mid MM, T)$ ---------- (6)

$P(Z_1, Z_2 \mid FF, T)$, ---------- (7)

and for non-targets:

$P(Z_1, Z_2 \mid MM, N) = P(Z_1 \mid M)\, P(Z_2 \mid M)$ ---------- (8)

$P(Z_1, Z_2 \mid FF, N) = P(Z_1 \mid F)\, P(Z_2 \mid F)$ ---------- (9)

The independence assumption in (8) and (9) holds when the model parameters are assumed known at the scoring time. In a more fully Bayesian treatment, the uncertainty in the estimates of the model parameters is taken into account during scoring.

If the continuous hidden variables have normal distributions, all these likelihoods can be computed in closed form and, if heavy-tailed distributions are considered then they can be approximated. The above mentioned likelihoods can be expressed by the following likelihood ratios:

$$R_M = \frac{P(Z_1, Z_2 \mid MM, T)}{P(Z_1 \mid M)\, P(Z_2 \mid M)} \quad \text{---------- (10)}$$

$$R_F = \frac{P(Z_1, Z_2 \mid FF, T)}{P(Z_1 \mid F)\, P(Z_2 \mid F)} \quad \text{---------- (11)}$$

$$G_1 = \frac{P(Z_1 \mid M)}{P(Z_1 \mid F)} \quad \text{---------- (12)}$$

Where log R_M and log R_F are gender dependent speaker verification scores for the male and female, respectively. log G_1 can be used as a gender discrimination score.

Likelihood Ratio Computation

The gender-independent likelihood ratio R can be obtained by marginalizing over the gender variables as [1]:

$$R = \frac{P(Z_1, Z_2 \mid T)}{P(Z_1 \mid M)\, P(Z_2 \mid M)} \quad \text{---------- (13)}$$

$$= \frac{P_M P(Z_1, Z_2 \mid MM, T) + P_F P(Z_1, Z_2 \mid FF, T)}{\Sigma_{g1,g2} Q_{g1,g2} P(Z_1 \mid g_1)\, P(Z_2 \mid g_2)}$$

Applying eq. (10) to eq. (12) in eq. (13), R can be expressed in terms of the likelihood ratios and priors as [4]:

$$R = \frac{P_M}{Q_{MM}} S_M R_M + \frac{P_F}{Q_{FF}} S_F R_F \quad \text{---------- (14)}$$

Where

$$S_M = \frac{Q_{MM} G_1 G_2}{Q_{MM} G_1 G_2 + Q_{MF} G_1 + Q_{MF} G_2 + Q_{FF}}$$

$$S_F = \frac{Q_{FF}}{Q_{MM} G_1 G_2 + Q_{MF} G_1 + Q_{FM} G_2 + Q_{FF}}$$

3. Case Study

The main aim of this paper is speaker verification, which consists of comparing a speech signal from an unknown speaker to a database of known speaker. The system can recognize the speaker, which has been trained with a number of speakers.

Providing robust speaker verification for applications with access to only short speech utterance is challenge. Problems of very short utterance with factor analysis approaches will be investigated in future. New wavelet design to handle speech veriability or spontaneous speech is required to developed in coming days. Realtime feature extraction is also a challenging task for future.

Some of the application of speaker verification are Speech to Speech translation for a pair of Indian languages, Speech enabled Office Suite, Voice dialing and Voice controlled answering machine, Smart Home, For disabled persons, Voice Activated Robots.

4. Conclusion

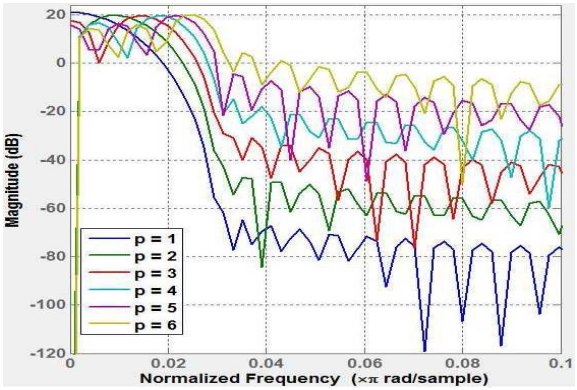


Figure 9: Frequency domain plot of six (M = 6), p is the taper index.Attenuation in the side-lobes decreases for higher order tapers

Figures 10 and11 provide a comparison of the multi-taper spectral estimates when uniform & non-uniform weights are applied, respectively. Adaptive weight in the Thomson multi taper method, in the context of speaker verification comparison suggest that non-uniform weights should be preferred.

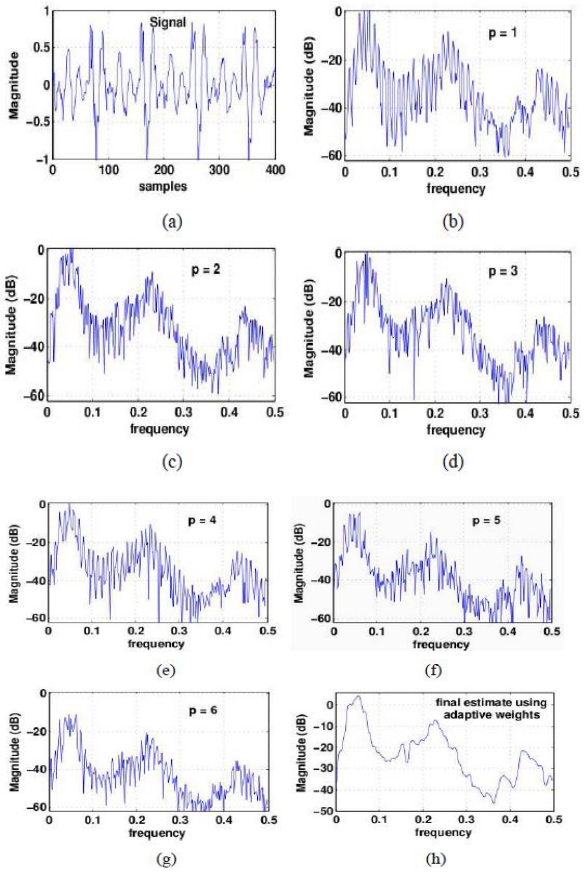
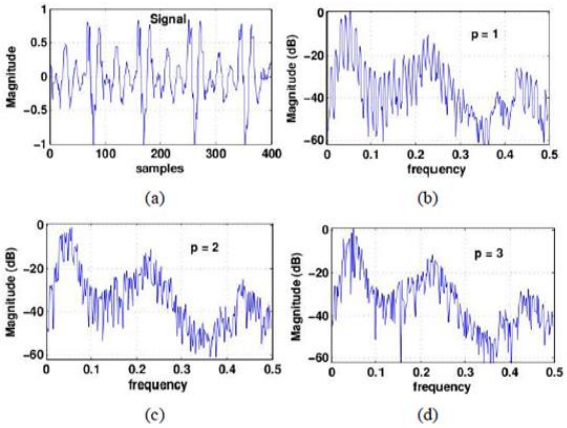


Figure 10: Multi-taper spectral estimates when adaptive weights are applied to the individual estimates (b)-(g) to get the final estimate (h) of a 25 msec duration speech signal (a)



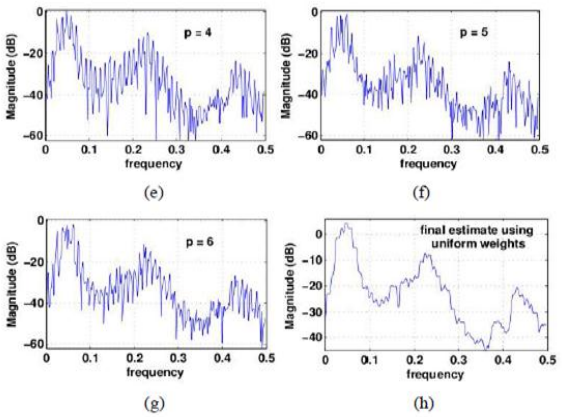


Figure 11: Multi-taper spectral estimates when uniform weights (1/M) are applied to the individual estimates (b)-(g) to get the final estimate (h) of a 25 msec duration speech signal (a)

Only the first taper (p = 1) in the multi-taper method produces a central peak at the harmonic frequency of the component while the other tapers (p > 1) produce spectral peaks that are shifted slightly up or down in frequency. The information lost at the extremes of the first taper is included and indeed emphasized in the subsequent tapers. attenuation in the side-lobes decreases with each taper in the sequence, i.e., spectral leakage increases for the higher-order tapers. If uniform weights are applied to get the final spectrum estimate, the energy loss at higher-order tapers will be high. In order to compensate for this increased energy loss, a weighted average (using non-uniform weights) is used instead of simply averaging the individual estimates. The weights are changed adaptively to optimize the bias/variance trade-off of the estimator. Multi-taper spectrum estimation approaches for low-variance MFCC and PLP feature computation and compared their performances, in the context of i-vector speaker verification, against the conventional single-taper (Hamming window) technique. In a Thomson multi-taper method, use of non-uniform weights, specifically adaptive weights, can bring improvement in speaker verification. Multi-taper method of MFCC and PLP feature extraction using i-vector is a viable candidate for replacing the baseline MFCC and PLP features.

References

[1] Md Jahangir Alama, TomiKinnunenc, Patrick Kennyb, Pierre Ouelletb, Douglas O'Shaughnessya, 2013. Multitaper MFCC and PLP features for speaker verification using i-vectors. Speech Communication 55(2):237-251.

[2] Alam, J.,CRIM, Montreal, QC, Canada, Kinnunen, T., Kenny, P., Ouellet, P. 2011. Multi-taper MFCC features for speaker verification using I-vectors. Automatic Speech Recognition and Understanding (ASRU), 547 - 552.

[3] M. Senoussaoui, P. Kenny, N. Brummer, E. de Villiers, and P. Dumouchel, 2011. Mixture of PLDA models in I-vector space for gender independent speaker recognition. Proc. of Interspeech, Florence, Italy, 25-28.

[4] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, 2011. Front-end factor analysis for speaker verification. IEEE Trans. on Audio, Speech and Language Processing, 19(4), 788-798.

[5] J. Sandberg, M. Hansson-Sandsten, T. Kinnunen, R. Saeidi, P. Flandrin, and P. Borgnat, 2010. Multitaper estimation of frequency-warped cepstra with application to speaker verification. IEEE Signal Processing Letters, 17(4), 343–346.

[6] Y. Hu and P. Loizou, 2004. Speech enhancement based on wavelet thresholding the multitaper spectrum. IEEE Trans. On Speech and Audio Proc., 12(1), 59-67.

[7] S. Davis and P. Mermelstein, 1980. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Trans. Acoustics, Speech, and Signal Processing, 28(2), 357–366.

[8] V.K. Khanna, "Digital Signal Processing", S. Chand & Company Ltd, 2nd Edition 2009.

Rajani P.K. is an Assistant Professor in E &TC Department in PimpriChinchwad College of Engineering,University of Pune, India. She is been more than 10 Yeats of teaching as well as industrial experience. She has published various international & national papers and jounals in various conference. Her Areas of interest is Digital Signal Processing and VLSI Design.

Namrata Kalkar received her B.E. degree from University of Amravati in 2008. She is currently pursuing M.E in VLSI and Embedded Systems from P.C.C.O.E, Pune University. She has 1 years of teaching experience.

Snehal Sawarkar received her B.Tech. degree from YCM University Nashik in 2010. She is currently pursuing M.E in VLSI and Embedded Systems from P.C.C.O.E, Pune University.