M. Tech. Credit Seminar Report, Electronic Systems Group, EE Dept, IIT Bombay submitted Nov 03

FEATURES AND TECHNIQUES FOR SPEAKER RECOGNITION

S. K. Singh (Roll No. 03307409)
Supervisor: Prof P. C. Pandey

Abstract

This paper aims at providing a brief overview into the area of speaker recognition. Speaker recognition can be classified into text dependent and the text independent methods. The features of speech signal that are being used (or have been used) for speaker recognition are presented briefly in this paper. The commonly used techniques of pattern matching for the decision making process in speaker recognition have been discussed and an example of a speech recognition system base on phoneme analysis using the harmonic features of speech is presented in the end of the paper.

1. Introduction

The speech signal contains many levels of information. Primarily a message is conveyed via the spoken words. At other levels, speech conveys the information about the language being spoken, the emotion, gender, and the identity of the speaker. The automatic recognition of speaker and speech recognition are very closely related. While speech recognition sets its goals at recognizing the spoken words in speech, the aim of automatic speaker recognition is to identity the speaker by extraction, characterization and recognition of the information contained in the speech signal.

The applications of speaker recognition technology are quite varied and continually growing. This technique makes it possible to use the speaker's voice for verification of their identity and thereafter enable the control access to services such as voice dialing and voice mail, tele-banking, telephone shopping, database access related services, information services, security control for confidential information areas, forensic applications, and remote access to computers. Speaker recognition technology is expected to create a host of new services that will make our daily lives more convenient.

Speaker recognition is a commonly used biometric today in most of the commercialization that has taken place for control of access to information services or user accounts on computers. Speaker recognition offers the ability to replace or augment the personal identification numbers and passwords with something that cannot be stolen or lost. There are two main factors [Rey02], that make speaker recognition a compelling biometric; (1) Speech is natural signal to produce that is not considered threatening by the users to provide, and (2) the telephone system provides a familiar network of sensors for obtaining and delivering the speech signal.

The usage of voice as the main and the only way of providing speakers identity has certain associated problems. The characteristics of a speakers voice can be corrupted by the characteristics of the communication channel or by the background noise. Therefore, speaker recognition systems should be capable of accepting wide range of variations of the users voice. The capability would allow other speakers with similar voice characteristics to be accepted by the system. For successful speaker recognition, understanding of the principles of human speaker recognition is essential and therefore speaker recognition should include a close study of clues that are used by humans in recognizing the speaker. Finding the stable features of voice is therefore the most important task for speaker recognition. The process of human speech generation along with various features and techniques that have been identified and developed for the purpose of speaker recognition are presented in the following sections.

2. Classification of Speaker Recognition Methods

The problem of speaker recognition can be divided into two major sub problems: s*peaker identification* and s*peaker verification*. Speaker identification can be thought of, as the task of determining who is talking from a set of known voices of speakers. It is the process of determining who has provided a given utterance based on the information contained in speech waves. The unknown voice comes from a fixed set of known speakers, thus the task is referred to as closed set identification. Speaker Verification on the other hand is the process of accepting or rejecting the speaker claiming to be the actual one. Since it is assumed that imposters (those who fake as valid users) are not known to the system, this is referred to as the open set task. Adding a none of the above option to the closed set identification task would enable merging of the two tasks, and it is called open set identification. "Error that can occur in speaker identification is the false identification of speaker and the errors in speaker verification can be classified into the following two categories: (1) false rejections: a true speaker is rejected as an imposter, and (2) False acceptances: a false speaker is accepted as a true one" [Imp94].

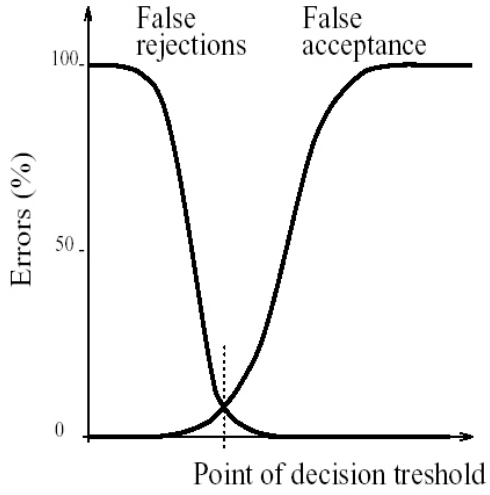

Figure 1. Decision threshold for false rejections and false acceptances. Reproduced from [Imp94]

In most systems for speaker recognition, a distance towards stored speaker's template is computed and is compared with predetermined threshold. If the computed distance is below the threshold the speaker is verified, otherwise speaker is rejected as an imposter. The decision threshold is located at the point where the probabilities of both the errors are equal as shown in the Figure 1.

Speaker recognition methods can also be divide into *text – dependent* and *text – independent* methods. In case of text – dependent methods a speaker is required to utter a predetermined set of words or sentences (e.g. a password). Features of voice are extracted from the same utterance. In case of text – independent methods, there is no predetermined set of words or sentences and the speaker's may not even be aware that they are being tested. Both the text – dependent and independent methods share a problem. These systems can be deceived because some one who plays back a recorded voice of a registered speaker saying the key words or sentences can be accepted as the registered speaker. Even the use of pre determined set of words or digits that are randomly chosen every time can be reproduced in the requested order by an advanced electronic recording equipment. Therefore a text – prompted (machine driven – text – dependent) speaker recognition system could be considered.

With the merger of speaker and speech recognition systems and improvement in speech recognition accuracy, the distinction between text – dependent and independent applications will eventually decrease. The text – dependent speaker recognition is the most commercially viable and useful technology, although there has been much research conducted on both the tasks. However, due to the possibilities offered, more attention is being paid to the text – independent methods of speaker recognition irrespective of their complexity.

3. The Human Voice

The origin of differences in voice of different speakers lies in the construction of their articulatory organs, such as the length of the vocal tract, characteristics of the vocal chord and the differences in their speaking habits. An adult vocal tract is approximately 17 cm long [Cam97] and is considered as part of the speech production organs above the vocal folds (earlier called as the vocal chords). As shown in Figure 2, the speech production organs includes the laryngeal pharynx (below the epiglottis), oral pharynx (behind the tongue, between the epiglottis and vellum), oral cavity (forward of the velum and bounded by the lips, tongue, and palate), nasal pharynx (above the velum, rear end of nasal cavity) and the nasal cavity (above the palate and extending from the pharynx to the nostrils). The larynx comprises of the vocal folds, the top of the cricoid cartilage, the arytenoids cartilages and the thyroid cartilage. The area between the vocal folds is called the glottis.

The resonance of the vocal tract alters the spectrum of the acoustic as it passes through the vocal tract. Vocal tract resonances are called formants. Therefore the vocal tract shape can be estimated from the spectral shape (e.g., formant location and spectral tilt) of the voice signal. Speaker recognition systems use features generally derived only from the vocal tract. The excitation source of the human vocal also contains speaker specific information. The excitation is generated by the airflow from the lungs, which thereafter passes through the trachea and then through the vocal folds. The excitation is classified as phonation, whispering, frication, compression, vibration or a combination of these. *Phonation* excitation is caused when airflow is modulated by the vocal folds. When the vocal folds are closed, pressure builds up underneath them until they blow apart. The folds are drawn back together again by their tension, elasticity and the Bernoulli effect. The oscillation of vocal folds causes pulsed stream excitation of the

vocal tract. The frequency of oscillation is called the fundamental frequency and it depends upon the length, mass and the tension of the vocal folds. The fundamental frequency therefore is another distinguishing characteristic for a given speaker.

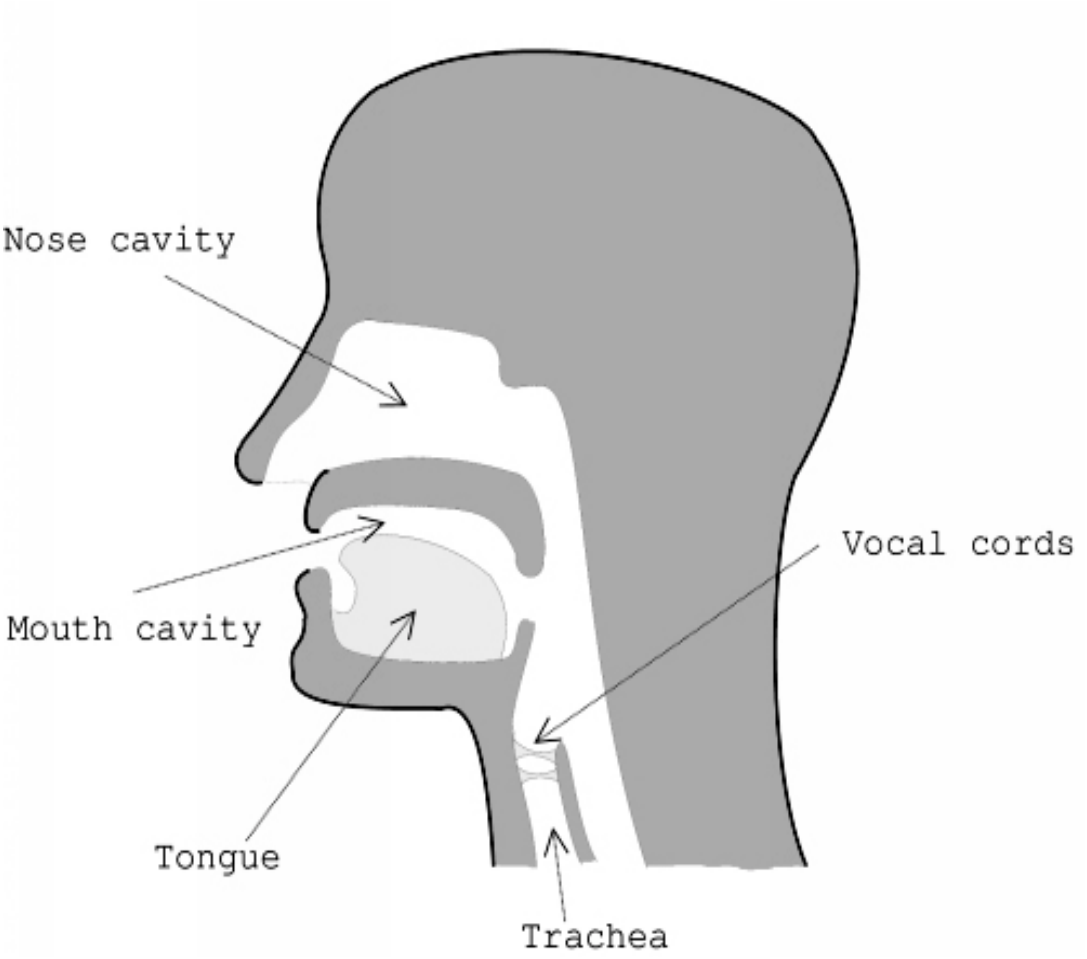

Figure 2. Human speech production organs. Figure downloaded from [Ttp03]

Whispered excitation is caused by the flow of air rushing through a small triangular opening between the arytenoids cartilages at the rear of the nearly closed vocal folds. A turbulent airflow results after this, which has a wide band noise characteristic. Frication excitation is caused due to the constrictions in the vocal tract. The shape of the broadband noise excitation depends upon the place, shape, and degree of constriction determine. The spectral concentration generally increases in frequency when the constriction moves forward. Sounds that are generated by friction are called fricatives. Frication can occur with or without phonation. *Compression* excitation is produced from release of a completely closed and pressurized vocal tract. This results in a silence (in the pressure accumulation phase) followed by a short noise burst. If the release is sudden, a *stop* or *plosive* is generated. If the release is gradual, an *affricate* is formed. Vibration excitation is a result of air being forced through a closure other than the vocal folds, especially at the tongue.

Speech produced by phonated excitation is called *voiced*, speech produced by phonated excitation plus frication is called *mixed voice* and speech produced by other types of excitation is called *unvoiced*. Due to the differences in the manner of production it is reasonable to expect some speech models to be more accurate for certain classes of excitation than the others. Unlike phonation and whispering the places of frication, compression and vibration excitation are actually inside the vocal tract itself. This could cause difficulties for models that assume an excitation at the bottom end of the vocal tract. The respiratory system also plays a role in the resonance properties of the vocal system of an individual. When the vocal folds are in vibration, resonances occur above and below the folds. Sub glottal resonances are largely dependent upon the properties of the trachea, which is typically 12 cm long and 2 cm in diameter [Cam97], made up of rings of cartilage joined together by connective tissue joining the lungs and the larynx. Due to this physiological dependence, the sub glottal resonances possess speaker dependent properties. Other physiological speaker dependent properties include vital capacity (the maximum volume of air one can blow out after maximum intake), maximum phonation time (the maximum duration a syllable can be sustained), phonation quotient (ratio of vital capacity to maximum phonation time) and glottal airflow (amount of air going through vocal folds). Other aspects of speech production that could be useful for discriminating between speakers are learned characteristics, including speaking rate, prosodic effects and dialect [Cam97].

4. Acquisition of Speech Signal

The acoustic wave speech signal generated by humans can be converted into an analog signal using a microphone. An antialiasing filter is thereafter used to condition this signal and additional filtering is used to compensate for the channel impairments. The antialiasing filter band limits the speech signal to approximately the Nyquist rate (half the sampling rate) before sampling. The conditioned analog signal is then sampled by an analog – to – digital (A/D) converter in order to obtain a digital signal. The A/D converters in use today for speech signal applications have a resolution of 12 to 16 bits typically at 8000 to 20,000 samples per second [Cam97]. For allowing the use of a simple antialiasing filter and precise control of the fidelity of the sampled speech signal, over sampling of the analog speech signal is used.

5. Basic Structure of Speaker Recognition System

Speaker recognition systems generally consist of three major units [Imp94] as shown in Figure 3. The input to the first stage or the front end processing system is the speech signal. Here the speech is digitized and subsequently the feature extraction takes place. There are no exclusive features that convey the speakers identity in the speech signal, however it is known from the source filter theory of speech production that the speech spectrum shape encodes in it the information about speakers vocal tract shape via formants and glottal source via pitch harmonics. Therefore some form or the other of the spectral based features is used in most of the speaker recognition systems. The final process in the front end processing stage is some form of channel compensation. Different input devices (e.g. different telephone handsets) impose different spectral characteristics on the speech signal, such as band limiting and shaping. Therefore channel compensation is done for removal of these unwanted effects. Most commonly some form of linear channel compensation, such as long and short-term cepstral mean subtraction are applied to features. The basic fundamental of spectral subtraction is that the power spectrum of speech signal corrupted by additive noise is equal to the sum of the signal power spectrum and noise

power spectrum. Power spectrum of the noisy signal is computed and from this spectrum an estimate of noise power spectrum is subtracted.

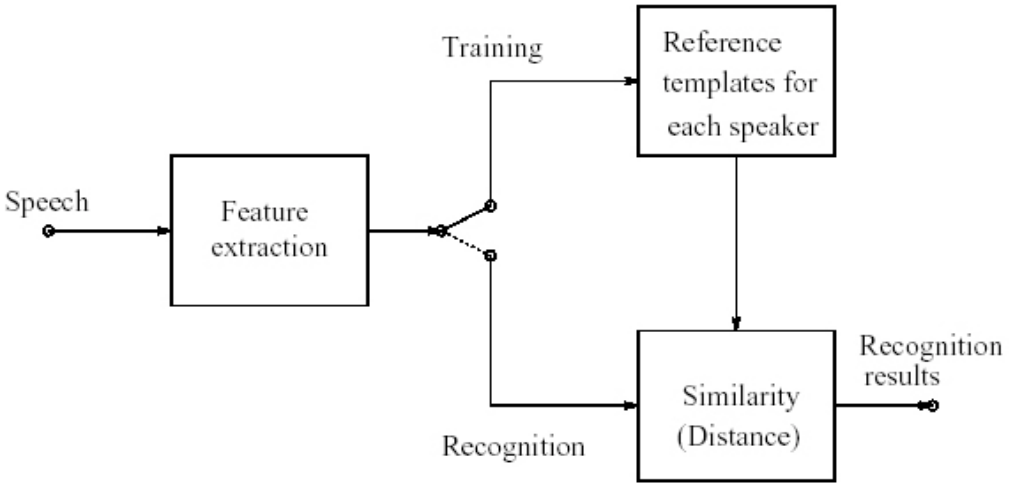

Figure 3. Structure of speaker recognition system. Reproduced from [Imp94].

The process of speaker recognition consists of the training phase and the recognition phase. In the training phase, the features of a speaker's speech signal are stored as reference features. The feature vectors of speech are used to create a speaker's model. The numbers of reference templates that are required for efficient speaker recognition depend upon the kind of features or techniques that the system uses for recognizing the speaker. In the recognition phase, features similar to the ones that are used in the reference template are extracted from an input utterance of the speaker whose identity is required to be determined. The recognition decision depends upon the computed distance between the reference template and the template devised from the input utterance. In speaker identification, the distance between an input utterance and all of the available reference templates is computed. The template of the registered user, whose distance with the input utterance template is the smallest, is finally selected as the speaker of the input utterance. In case of speaker verification the distance is computed only between the input utterance and the reference template of the claimed speaker. If the distance is smaller than the predetermined threshold, the speaker is accepted other the speaker is rejected as an imposter.

6. Features for Speaker Recognition

The speech signal can be represented by a sequence of feature vectors in order to application of mathematical tools without the loss of generality. Most of these features are also used for speaker dependent speech recognition systems. In practical real life systems, several of these features are used in combinations. Some of the desirable properties for feature sets [Faq00] are as follows:

I. They should preserve or highlight information and variation in the speech that is relevant to the basis being used for the speech recognition and at the same time minimize or eliminate any variation irrelevant to that task.

II. Feature space should be relatively compact in order to enable easier learning of models from finite amounts of data.

III. A feature representation that can be used without much consideration in most circumstances should be used.

IV. The process of feature calculation should be computationally inexpensive. Processing delay (i.e. how much of the 'future' of the signal you have to know before you can emit the features) is a significant factor in some settings, such as real-time recognition.

The subsequent sections briefly describe certain features that are used (or have been used) in speaker recognition systems.

6.1. Frequency Band Analysis

Filter banks were initially used to gather information about the spectral structure of signal. The filter banks consist of number of filters where each filter covers one group of frequencies. Bandwidths of filters could be chosen to be equal, Logarithmic or may correspond to certain critical intervals. The output of such filter bank offers largely depends upon the number of the filters being used, which normally varies from 10 – 20 and thus this technique represents an approximate representation of the actual spectrum. The output of the filter bank is sampled (usually 100 Hz) and the samples of the output indicate the amplitude of the frequencies from a particular bandwidth. The output is thus used as the feature vector for speaker recognition.

6.2. Formant Frequencies

Periodic excitation is seen in the spectrum of certain sounds, especially vowels. The speech organs form certain shapes to produce the vowel sound and therefore regions of resonance and anti resonance are formed in the vocal tract. Location of these resonance's in the frequency spectrum depends on the form and shape of the vocal tract. Since the physical structure of the speech organs is a characteristic of each speaker, differences between speakers can also be found in the position of their formant frequencies. The resonance's heavily affect the overall spectrum shape and are referred to as *formants*. A few of these formant frequencies can be sampled at an appropriate rate and used for speaker recognition. These features are normally used in combination with other features.

6.3. Pitch Contours

The variations of the fundamental frequency (pitch) during the duration of the utterance if followed, would provide the contour, which can be used as a feature for speech recognition. The speech utterance is normalized and the contour is determined. The normalization of the speech utterance is required because the accurate time alignment of utterances is crucial; else the same speaker utterances could be interpreted as utterances from two different speakers. The contour is divided into a set of segments and the measured pitch values are averaged over the whole segment. The vector that contains the average values of pitch of all segments is thereafter used as a feature for speaker recognition.

6.4. Coarticulation

Coarticulation is a phenomenon where a feature of a phonemic unit is achieved in the articulators well in advance of the time it is needed for that phonemic unit. Variation of the physical form of the speech organs causes the variation in the sounds that they produce. The process of coarticulation in which, the speech organs prepare to produce a new sound while

transiting from one sound to another is characteristic of a speaker. This is due to the following reasons: the construction and shape of the vocal tract, and the motorical abilities of the speaker to produce the sequences of speech. Therefore for speaker recognition using this feature, the points in the speech signal where coarticulation takes place are spectrographically analyzed.

6.5. Features derived from Short term processing

The following features of the short - term processing of the speech can be applied [Imp94]: short - term autocorrelation, average magnitude difference function, zero crossing measure, short - term power and energy measures, and short - term Fourier analysis. The short term processing techniques provide signals in the following form: -

$$Q(n) = \sum_{m=-\infty}^{\infty} T[s(m)]w(n-m) \qquad \ldots. (1)$$

T[s (m)] is a transformation, which is applied to the speech signal and the signal is thereafter weighted by a window *w (n)*. The summation of *T[s (n)]* convolved with *w (n)* represents certain property of the signal that is averaged over the window duration.

6.5.1 Short Time Average Energy and Magnitude

The output in the equation (1) will be representing short time energy or amplitude if the transformation *T* is squaring or absolute magnitude operation. The energy indicates high amplitudes as the signal is squared for calculating *Q (n)*. Such techniques enable the segmentation of speech into smaller phonetic units e.g. phonemes or syllables. There is a large variation in the amplitude between the voiced and the unvoiced segments. Also, the variation between phonemes with different manners of articulation is small. This feature permits speech segmentation based on energy *Q (n)* in automatic recognition systems.

6.5.2 Short Time Average Zero Crossing Rate

A zero crossing is said to have occurred in a signal when its waveform crosses the time axis or changes its algebraic sign. For a discrete time signal with zero crossing rate (ZCR) in zero crossings/sample and a sampling frequency of *Fs*, the frequency *Fo* is given as

$$Fo = (ZCR * Fs)/2 \qquad \ldots. (2)$$

The speech signal contains most of its energy in voiced signals at low frequencies. For unvoiced sounds, the broadband noise excitation takes place at higher frequencies due to the short length of the vocal tract. Therefore a high and a low ZCR relates to unvoiced and voiced speech respectively.

6.5.3 Short Time Autocorrelation

The autocorrelation function for a discrete time signal is given as [Sha01]: -

$$\Phi(k) = \sum_{m=-\infty}^{\infty} s(m)y(m-k) \qquad \ldots. (3)$$

This function measures the similarity of two signals *s (n)* and *y (n)*, by summing the product of a signal sample and a delayed sample of another signal. The short time autocorrelation

function is obtained by windowing *s (n)* and applying the autocorrelation given by equation (3), which results in

$$R_n(k) = \sum_{m=-\infty}^{\infty} s(m)w(n-m)s(m-k)w(n-m+k) \qquad \ldots. (4)$$

This short time auto correlation function provides information about the harmonic and formant amplitudes of *s (n)* and also indicates its periodicity. Thus pitch estimation and voiced/unvoiced speech detection can be carried out using this feature.

6.6. Linear Prediction Features

The basic idea of linear prediction is that a speech sample *s (n)* related to excitation *u (n)* can be predicted (approximated) by a linear combination of the past *P* speech samples, as the following equation shows [Rsc78]: -

$$s(n) = \sum_{k=1}^{p} a_k s(n-k) + Gu(n) \qquad \ldots. (5)$$

Here *G* is the gain parameter and a_k are the prediction coefficients. The steady state system function for speech production model is given as below: -

$$H(z) = \frac{S(z)}{U(z)} = \frac{G}{1 - \sum_{k=1}^{p} a_k z^{-1}} \qquad \ldots. (6)$$

The linear prediction model signal with the prediction coefficient α_k is given as below: -

$$\hat{s}(n) = \sum_{k=1}^{p} \alpha_k s(n-k) \qquad \ldots. (7)$$

The prediction error *e (n)* is difference between the actual and the approximated speech sample

$$e(n) = s(n) - \hat{s}(n) = s(n) - \sum_{k=1}^{p} \alpha_k s(n-k) \qquad \ldots. (8)$$

e (n) can be viewed as the output of a system with the following transfer function

$$A(z) = \frac{E(z)}{S(z)} = 1 - \sum_{k=1}^{p} a_k z^{-1} \qquad \ldots. (9)$$

Comparing equations (5) and (8) we see that if $a_k = \alpha_k$, then

$$e(n) = Gu(n) \qquad \ldots. (10)$$

and

$$s(n) = \sum_{k=1}^{p} \alpha_k s(n-k) + e(n) \quad \text{.... (11)}$$

Thus the prediction error filter *A (z)* is the inverse filter for system *H(z)* of equation (6)

$$H(z) = \frac{G}{A(z)} = \frac{G}{1 - \sum_{k=1}^{p} a_k z^{-1}} \quad \text{.... (12)}$$

The prediction coefficients of the inverse filter are determined by linear prediction analysis such that the error *e(n)* is minimized. The mean square error can be expressed as the sum of the squared differences between the predicted samples and the actual ones over a finite interval.

$$\varepsilon = \sum_{k=1}^{N} \varepsilon_k = \sum_{k=1}^{N} \left[s(n) - \hat{s}(n) \right]^2 \quad \text{.... (13)}$$

$$= \sum_{N} \left[s(n) - \sum_{k=1}^{p} \alpha_k s(n-k) \right]^2 \quad \text{.... (14)}$$

Where, ε = sum of the squared errors,
ε_k = Squared error,
N = interval length,

By the minimization of the sum of the squared error the best linear prediction coefficients can be determined. The minimization leads to a flat (band limited white) response. Thus *A (z)* is also called whitening filter. If the voiced speech signal suits the linear prediction based speech model then the error signal is a good approximation of the excitation source. The error signal would look like a train of impulses that repeats the rate of vocal folds vibration. The maximum prediction errors will occur at the vocal fold vibration rate thus indicating the pitch period. The linear prediction coefficients can be directly used in digital filters as multipliers or can be stored as templates in speech recognizers.

6.7. Harmonic Features

The harmonic decomposition of the high-resolution spectral line estimate of speech signal results in the harmonic features. The line spectral pairs represent the variations in the glottis and the vocal tract of a speaker, which are transformed into frequency domain. The feature vector of harmonic features contains the fundamental frequency followed by amplitudes of several harmonic components. These features can be produced only on voiced segments of speech and the long vowels and nasals were found to be most speaker specific [Hor95].

7. Template Matching

The features of the speech signal are in the form of N – dimensional feature vector. For a segmented signal that is divided into M segments, M vectors are determined producing the $M \times N$ feature matrix. The $M \times N$ matrix is created by extracting features from the utterances of the speaker for selected words or sentences during the training phase. After extraction of the feature vectors from the speech signal, matching of the templates is required to be carried out for speaker recognition. This process could either be manual (comparison of spectrograms visually) or automatic. In automatic matching of templates, speaker models are constructed from the extracted features. There after a speaker is authenticated by comparison of the incoming speech signal with the stored model of the claimed user. The speaker models are of two types: template models and stochastic models.

7.1 Template Models

The simplest template model has a single template $\boldsymbol{x}$, which is the model for a speech segment. The match score between the template $\boldsymbol{x}$ for the claimed speaker and an input feature vector $\boldsymbol{y}$ from an unknown user is given by $\boldsymbol{d(x, y)}$. The model for the claimed speaker could be the centroid (mean) of a set of $\boldsymbol{N}$ vectors obtaining in training phase [Cam97].

$$x = \frac{1}{N}\sum_{k=1}^{N} y_k \qquad \text{.... (15)}$$

The various distance measures between the vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ can be written as

$$d(x, y) = (x - y)^T W (x - y) \qquad \text{..... (16)}$$

Where, W is the weighting matrix. If W is an identity matrix, then all the elements of the vectors are equally treated and the distance is called *Euclidean*. If W is a positive – definite matrix that would allow desired weighting of the template features then, the distance is *Mahalanobis*.

7.1.1 Dynamic Time Warping (DTW)

The time alignment of different utterances is a serious problem for distance measures and a small shift would lead to incorrect identification. Dynamic time warping is an efficient method to solve this time alignment problem. This is the most popular method for speaking rate variability in template-based systems [Cam97]. The asymmetric match score $\boldsymbol{\beta}$ of comparison of an input frame $\boldsymbol{y}$ of M samples with the template sequence $\boldsymbol{x}$ is given as follows

$$\beta = \sum_{i=1}^{M} d(y_i, x_{j(i)}) \qquad \text{.... (17)}$$

The template indices $j(i)$ are given by the DTW algorithm. This algorithm performs a piece wise linear mapping of the time axis to align both the signals. The variation over time in the parameters corresponding to the dynamic configuration of the articulators and the vocal tract is taken into account in this method. Figure 4, shows the dynamic time warp of two energy signals. The warp path is a diagonal line for two identical signals and the warp has no effect. The

accumulated deviation from the dashed diagonal warp path is the Euclidean distance between two signals and the parallelogram surrounding the warp path acts as boundary conditions for preventing excessive warping.

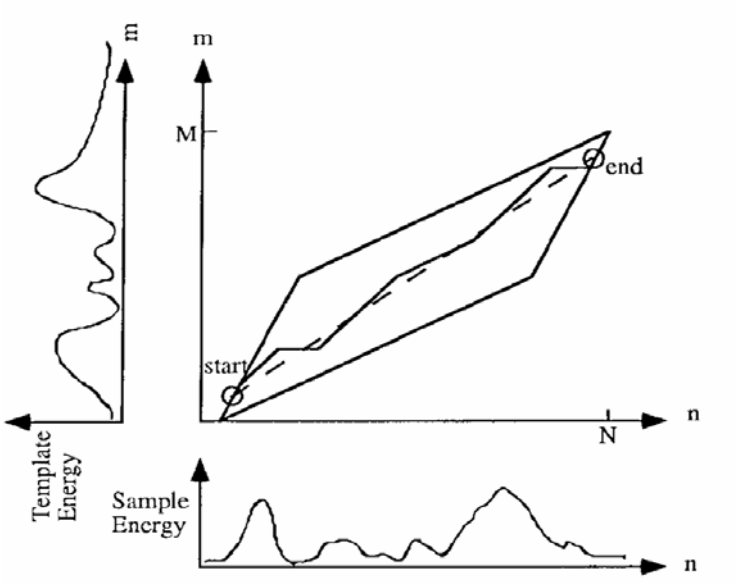

Figure 4. DTW of two energy signals. Reproduced from [Cam97]

7.1.2 VQ Source Modeling

This is another form of usually text dependent template model that uses multiple frames of speech. This model makes use of has a vector quantized codebook, which is generated for a speaker by using his/her training data. Standard clustering procedures are utilized for formulation of the codebook. These procedures average out the temporal information from the codebook and therefore the requirement of performing time alignment is eliminated. The pattern match score is the distance between the input vector and the minimum distance code word in the codebook. This method is simplified due to the lack of time warping but neglects the speaker dependent temporal information, which may be present in the prompted text [Cam97].

7.1.3 Nearest Neighbors

This method combines the strengths of the dynamic time warping and vector quantization methods. This method keeps all the data obtained from training phase and does not cluster data to obtain the codebook. Therefore it can make use of the temporal information that may be present in the prompted phrase. The distances between the input frames and the stored frames is used for computing the inter frame distance matrix. The nearest neighbor distance is the minimum distance between the input and the stored frames. The nearest neighbor distances for all input frames are averaged to arrive upon the matched score. These matched scores are thereafter combined to form an approximation of the likelihood ratio. This method is very memory intensive and is one of the most powerful methods [Cam97].

7.2. Stochastic Models – Hidden Markov Model

Stochastic models have been a lately which provide more flexibility and produce better matching score. In a stochastic model, the process of pattern matching is carried out by measuring the likelihood of a feature vector in a given speaker model. A stochastic model that is widely used for modeling of sequences is the Hidden Markov Model [Cam97]. This technique efficiently models the statistical variations of the features and provides a statistical representation of the manner in which a speaker produces sounds.

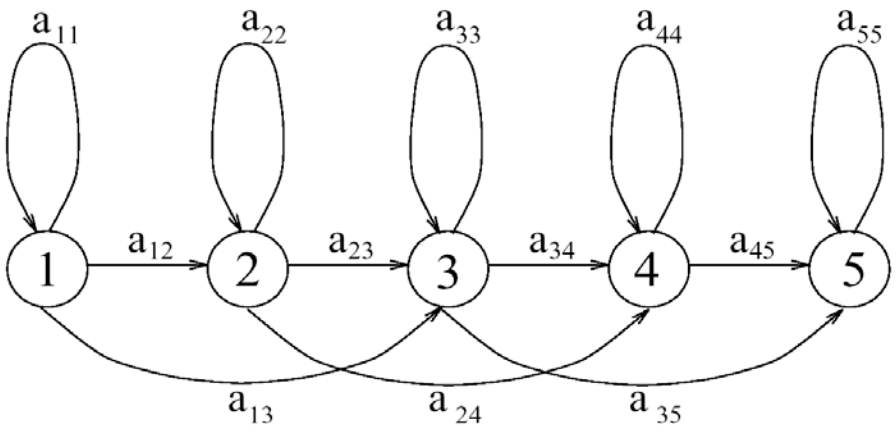

Figure 5. Five – state Markov model. Reproduced from [Imp94].

A Hidden Markov Model (HMM) consists of a set of transitions between a set of states. Two sets of probabilities are defined for each transition: a transition probability and the output probability density function. The output probability density function is the probability of emitting each of the output symbols from a finite vocabulary [Imp94]. As shown in Fig. 5, the transitions are allowed to the next right state or the same state, thus the model is named *left – to right model* and a_{ij} are the probabilities of transition to other states. The HMM parameters are generated from the speech during the training phase and for verification, the likely hood of the input feature sequence is computed with respect to the speakers HMMs. In case of finite vocabulary being used for speaker recognition, each word is modeled using multiple state *left – to right* HMMs. Therefore in case of large vocabulary, larger number of models are required.

8. Example Of Speaker Recognition System

Figure 6, and 7 present the block schematics of text dependent speaker recognition systems. In the system of Fig. 6, features derived from the LPC coefficients are used for speaker modeling and one of the LPC features extracted from a speech signal is utilized [Imp94]. The classification of templates is done by dynamic time warping and thereafter the weighting of reference templates is carried out. The system in the figure 7 employs the statistical features of spectral parameters extracted from a word utterance. Speaker recognition decision is arrived upon by computing and

comparing the weighted distances between templates. “The text depending speaker recognition systems achieve as much as 98% to 100% accuracy with more than 20 speakers” [Imp94], but the performances of the text independent speaker recognizers are far behind. Long-time averaging of speech spectrum is the most common approach to text independent speaker recognition, where the length of the speech utterance may be more than 10 seconds. Averaging the values of the spectrum features of the whole utterance does speaker recognition.

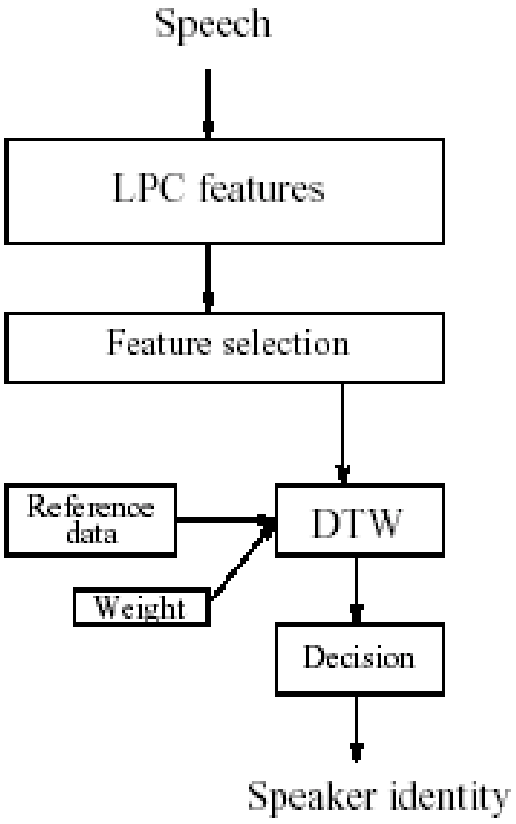

Figure 6: Block diagram of speaker recognition system utilizing LPC features and DTW. Reproduced from [Imp94].

These method posses many disadvantages such as the larger length of the required speech signal and exhibit low performances especially when the number of speakers increases. Improved systems of text independent speaker recognition incorporate phoneme recognition methods. In the phoneme recognition method, phonemes are recognized for each segment of the speech by using the speaker independent phone reference templates. The distance between the recognized phoneme and the same phonemes of all the registered speakers is subsequently determined. This procedure is repeated for all the segments of a given utterance and an average value of the distances is computed for each speaker, which thereafter accomplishes the speaker recognition. Short segments of speech are also compared to the reference templates of each speaker and the smallest distance for each speaker is utilized for determining the average distances for each speaker. The speaker with minimum average distance is finally selected.

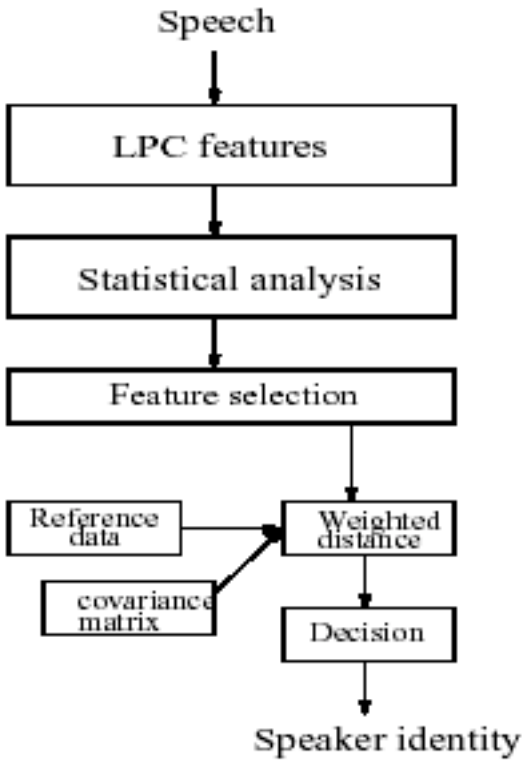

Figure 7: Block diagram of speaker recognition system utilizing LPC features and weighted distance measure. Reproduced from [Imp94].

9. Conclusion

There has been a considerable amount of development in the field of speech and speaker recognition. The techniques for speaker recognition are yet to be successfully used in practical systems as the recognition rate is drastically reduced due to many reasons such as the distortion in the channel and the recording conditions and speaker-generated variability. Therefore it is important to explore stable features that remain insensitive to variation of speakers voice over time and are robust against variation in voice quality due to colds or disguises. The problem of distortion in the channels and background noise also requires being resolved with better techniques.

Acknowledgement

I express deep gratitude to my guide, Prof. P. C. Pandey, for his invaluable support and guidance, which effectively contributed in successful completion of this seminar. He was instrumental in providing technical, moral and administrative support. It is a privilege to work under his guidance.

References

[Cam97] Joseph P. Campbell, ”Speaker recognition: A tutorial”, Proc. IEEE, vol. 85, pp. 1437-1462, September 1997.

[Faq00] ICSI Speech FAQ: 5.1, “What are features? What are their desirable properties?”, Answer by: dpwe - 2000-05-26, downloaded from **http://www.icsi.berkely.edu/local-cgi-bin/man-cgi-bin?ftrs-intro.html** on Tue Mar 25 20:52:32, 2003.

[Hor95] Bojan Imperl, Zdravko, and Bogomir Horval, “Use of Harmonic Features in Speaker recognition”, Laboratory for Digital Signal Processing, Faculty of Electrical Engineering and Comp. Sci., Smetanova 17, 2000 Maribor, Slovenia.

[Imp94] Bojan Imperl, “Speaker recognition techniques”, Laboratory for Digital Signal Processing, Faculty of Electrical Engineering and Comp. Sci., Smetanova 17, 2000 Maribor, Slovenia.

[Rey02] D. A. Reynolds, “An Overview of Automatic Speaker Recognition Technology”, Proc. IEEE, pp. 4072-4075, 2002.

[Rsc78] L. Rabiner and R. Schafer, *Digital Processing Speech Signals*, Englewood Cliffs, NJ: Prentice Hall, Inc., 1978.

[Sha01] D. O’Shaughnessy, *Speech Communications – Human and Machine*, Universities Press (India) Limited, 2001.

[Ttp03] Figure downloaded on Mon, Nov 10, 21:34:05, 2003 from http://lumumba.luc.ac.be/jori/thesis/onlinethesis/chapter4.html#fig5.)