Vector Quantization based Speaker Identification

Manjot Kaur Gill
Assistant Professor (CSE/IT)
Guru Nanak Dev Engg. College,
Ludhiana, Punjab

Reetkamal Kaur
Assistant Professor (CSE)
RIMT-MAEC,
Mandi Gobindgarh, Punjab

Jagdev Kaur
Student (M. Tech)
Guru Nanak Dev Engg.College,
Ludhiana, Punjab

ABSTRACT
The automatic speaker recognition technologies have developed into more and more important modern technologies required by many speech-aided applications. The main challenge for automatic speaker recognition is to deal with the variability of the environments and channels from where the speech was obtained. Speaker recognition system is a system which recognizes the speaker as opposed to what is being said by the speaker as in case of speech recognition. Speaker recognition technology makes it possible to the speaker's voice to control access to restricted services, for example, phone access to banking, database services, shopping or voice mail, and access to secure equipments. The main aim of this paper is speaker identification, which consists of comparing a speech signal from an unknown speaker to a database of known speakers. The methodology followed in this paper for Speaker identification is using Feature Extraction process and then Vector Quantization of extracted features is done using k-means algorithm. At last, the speaker is identified by comparing the data from a tested speaker to the database of each speaker and then measuring the difference.

Key Words: Cepstrum, K-means, Mel Scale, Speaker Identification, Vector Quantization.

1. INTRODUCTION
Speaker recognition refers to task of recognizing peoples by their voices. It is the process of automatically recognizing who is speaking on the basis of individual information included in speech waves. It is basically divided into speaker identification and speaker verification. A speaker identification system gets a test utterance as input. The task of the system is to find out which of the training speakers made the test utterance. So, the output of the system is the name of the training speaker or speaker ID, or possibly a rejection if the utterance has been made by an unknown person. Verification is the task of automatically determining if a person really is the person he or she claims to be [1]. This technology can be used as a biometric feature for verifying the identity of a person in applications like banking by telephone, voice dialling telephone shopping, information services, voice mail and security control for secret information areas. Speaker recognition technology is the most potential technology to create new services that will make our everyday lives more secured. Another important application of speaker recognition technology is for forensic purposes.

As speaker recognition attempts to identify the person who is speaking but in speech recognition is the main stress is on what a person is speaking and not who is speaking. Speech recognition is a process of converting an acoustic signal, captured by a microphone or a telephone, to a set of words. The recognized words can be the final results in written format. They can also serve as the input to further linguistic processing in order to achieve speech understanding.

This paper takes speaker identification into consideration, which consists of mapping a speech signal from an unknown speaker to a database of known speakers, i.e. the system has been trained with a number of speakers which the system can recognize. There are two modes of operation that are related to the set of known voices. In the closed-set mode, the system assumes that the to-be-determined voice must come from the set of known voices. Otherwise, the system is in open-set mode. The closed-set speaker identification can be considered as a multiple-class classification problem. In open-set mode, the speakers that do not belong to the set of known voices are referred to as impostors. This task can be used for forensic applications, e.g., speech evidence can be used to recognize the perpetrator's identity among several known suspects. On the other hand, speaker verification is the process of rejecting or accepting the identity claim of a speaker. In most of the applications, voice is used as the key to confirm the identities of a speaker and is classified as speaker verification.

Speaker recognition systems are categorized into text-dependent and text-independent methods [9]. A text independent speaker recognition system does not have any information about the content of training and test utterances. On the contrary, a text dependent system relies on the restriction that the text that is said in training is identical to the test utterance. Text-dependent systems require the speaker to utter a specific phrase (pin-code, password etc.) while a text-independent method should catch the characteristics of the speech irrespective of the text spoken. Formerly text dependent methods were widely in application, but later text independent is in use. Speaker identification has been done successfully using Vector Quantization (VQ). This technique consists of extracting a small number of representative feature vectors as an efficient means of characterizing the speaker specific features. Using training data, these features are clustered to form a speaker-specific codebook. In the recognition stage, the test data is compared to the codebook of each reference speaker and a measure of the difference is used to make the recognition decision. Speaker verification is somewhat complex as compared to speaker identification. In speaker identification, the procedure that is followed in this paper consists of giving input speech signal to feature extraction stage which extracts the useful part from the speech signal like pitch variations and also different styles of speaking same word by different speakers are taken care by this stage. Some speakers put stress on specific part of a word and others do not. Then the output is given to next matching stage which checks the similarity score of unknown speaker with the database of known speakers and accordingly the decision is made and speaker name or speaker ID is the output of whole process. The whole process of speaker identification is depicted in figure1.

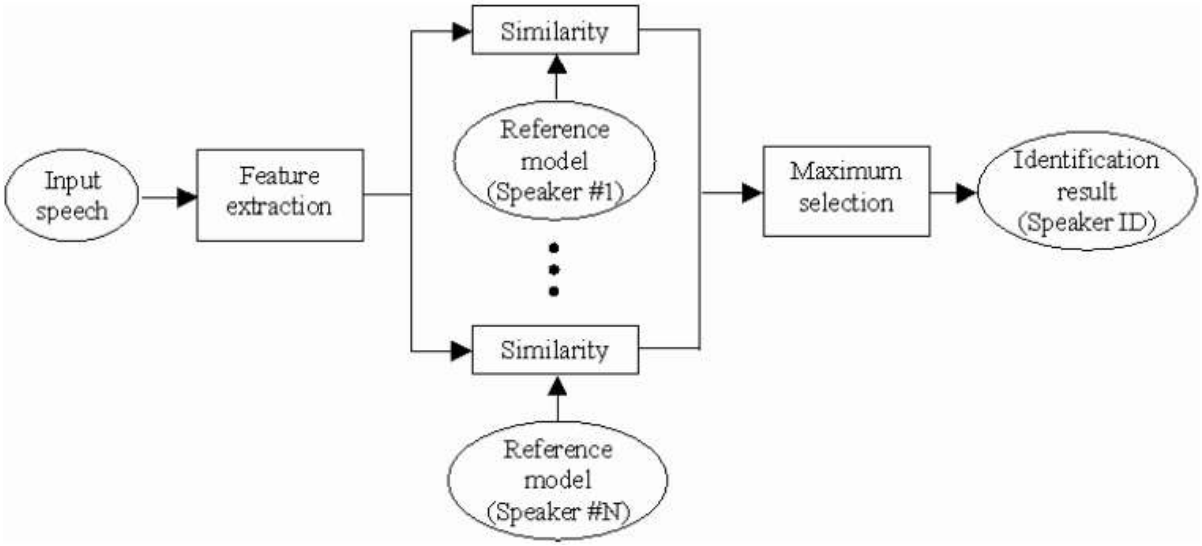

Figure 1: Conceptual presentation of Speaker Identification [1].

The Vector Quantization in this paper is presented utilizing Mel Frequency Cepstral Coefficient and a simple clustering scheme using the k-means algorithm [8].

2. SPEECH FEATURE EXTRACTION

The primary goal of feature extraction is to simplify recognition by summarizing the vast amount of speech data and obtaining the acoustic properties that define speaker individuality. The feature extractor converts the digital speech signal into a sequence of numerical descriptors, called feature vectors. The features provide a more stable, robust, and compact representation than the raw input signal. Feature extraction can be considered as a data reduction process that attempts to capture the essential characteristics of the speaker with a small data rate. This stage is often referred as speech processing front end. MFCC (Mel Frequency Cepstral Coefficients) is one of the most widely used feature extraction techniques. 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. In speaker identification the number of training and test vector needed for the classification problem grows exponential with the dimensions of the given input vector, so the need for feature extraction arises.

Feature extraction is a necessary operation for two main reasons. First, in order the statistical speaker models to be robust, the number of training samples must be large enough compared to the dimensionality of the measurements. The amount of needed training vectors grows exponentially with the dimensionality. This phenomenon is known as curse of dimensionality [3]. The second reason for performing feature extraction is the reduced computational complexity.

But extracted feature should meet some criteria while dealing with the speech signal [5]. Such as:

- discriminate between speakers while being tolerant of intra-speaker variabilities,
- easy to measure,
- stable over time,
- occur naturally and frequently in speech,
- change little from one speaking environment to another,
- not susceptible to mimicry.

For speech signals, it is known that the best features are based on spectral analysis. The reason for that is that the speech signal can be estimated with a linear superposition of sine-waves with different amplitudes and phases.

In this paper we are presenting the Mel Frequency Cepstral Coefficients (MFCC) technique to extract features from the speech signal and comparing the unknown speaker with the existent speaker in the database. The steps that are followed are shown in figure 2.

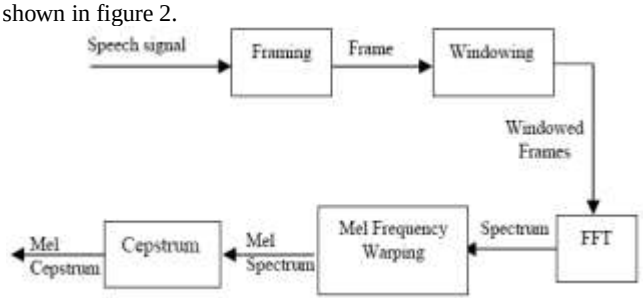

Figure 2: Feature Extraction Steps

2.1 Framing and Windowing

The speech signal is slowly varying over time (quasi-stationary) that is when the signal is examined over a short period of time (5-100msec), the signal is fairly stationary.

Therefore speech signals are often analyzed in short time segments, which are referred to as short-time spectral analysis. This practically means that the signal is blocked in frames of typically 20-30 msec. Adjacent frames typically overlap each other with 30-50%, this is done in order not to lose any information due to the windowing.

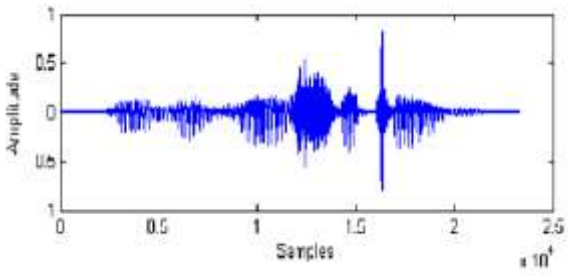

Figure 3: Wav format of speech signal

After the signal has been framed, each frame is multiplied with a window function $w(n)$ with length N, where N is the length of the frame [6]. Typically the Hamming window is used:

$$w(n)=0.54-0.46\cos(2\Pi n / N-1), \quad \text{where } 0 \le n \le N-1 \quad ...(i)$$

The windowing is done to avoid problems due to truncation of the signal as windowing helps in the smoothing of the signal.

2.2 Cepstrum

Cepstrum name was derived from the spectrum by reversing the first four letters of spectrum. Cepstrum is the Fourier Transform of the log with unwrapped phase of the Fourier Transform. The speech signal is composed of a quickly varying part $e(n)$(excitation sequence) convolved with a slowly varying part $\Theta(n)$ (vocal system impulse response) [2, 10]:

$$s(n)=e(n)* \Theta(n) \quad ...(ii)$$

The convolution makes it difficult to separate the two parts, therefore the cepstrum is introduced. The cepstrum is defined in the following way:

$$c_s(n)= \zeta^{-1}\{\log |\zeta\{s(n)\}|\} \quad ...(iii)$$

where, ζ is the Discrete Time Fourier Transform and ζ^{-1} and is the Inverse Discrete Time Fourier Transform. By moving the signal to the frequency domain, the convolution becomes a multiplication:

$$S(w)=E(w)\,\Theta(w) \quad ...(iv)$$

Further, by taking the logarithm of the spectral magnitude the multiplication becomes an addition:

$$\log|S(w)|=\log|E(w)\,\Theta(w)|$$
$$\log|S(w)|= \log|E(w)| + \log|\Theta(w)|$$
$$\log|S(w)|= Ce(w) + C_\Theta(w) \quad ...(v)$$

The Inverse Fourier Transform is linear and therefore it works individually on the two components:

$$c_s(n) = \zeta^{-1}(Ce(w) + C_\Theta(w))$$
$$c_s(n) = \zeta^{-1}\{Ce(w)\} + \zeta^{-1}\{C_\Theta(w)\}$$
$$c_s(n) = c_e(n)+c_\Theta(n) \quad ...(vi)$$

2.3 Mel Frequency Cepstral Coefficients (MFCC)

Mel frequency Cepstral Coefficients are coefficients that represent audio based on perception. This coefficient has a great success in speaker recognition applications. It is derived from the Fourier Transform of the audio clip. In this technique the frequency bands are positioned logarithmically, whereas in the Fourier Transform the frequency bands are not positioned logarithmically. As the frequency bands are positioned logarithmically in MFCC, it approximates the human system response more closely than any other system [12]. These coefficients allow better processing of data.

In the Mel Frequency Cepstral Coefficients the calculation of the Mel Cepstrum is same as the real Cepstrum except the Mel Cepstrum's frequency scale is warped to keep up a correspondence to the Mel scale. The Mel scale is mainly based on the study of observing the pitch or frequency perceived by the human. The scale is divided into the units mel. The mel scale is generally speaking a linear mapping below 1000 Hz and logarithmically spaced above. The mapping is usually done using an approximation (where *fmel* is the perceived frequency in mels):

$$fmel = 2595 * \log 10(1 + f/700) \quad ...(vii)$$

2.4 Delta Cepstrum

The cepstral coefficients provide a good representation of the local spectral properties of the framed speech. But, it is well known that a large amount of information resides in the transitions from one segment of speech to another. An improved representation can be obtained by extending the analysis to include information about the temporal cepstral derivative. Delta Cepstrum is used to catch the changes between the different frames. Delta Cepstrum defined as:

$$\Delta Cs(n;m) = 1\backslash 2\{Cs(n;m+1) - Cs(n;m-1)\}, \quad i=1,\ldots, Q \quad ...(viii)$$

The results of the feature extraction are a series of vectors characteristic of the time-varying spectral properties of the speech signal.

3. VECTOR QUANTIZATION

Speaker recognition is the task of comparing an unknown speaker with a set of known speakers in a database and finding the best matching speaker. Vector quantization (VQ) is a lossy data compression method based on the principle of block coding. Vector quantization can be thought of as a process of redundancy removal that makes the effective use of nonlinear dependency and dimensionality by compression of speech spectral parameters. Generally, the use of vector quantization results in a lower distortion than the use of scalar quantization at the same rate [4]. Vector Quantization is one of the preferred methods to map vast amount of vectors from a space to a predefined number of clusters each of which is defined by its central vectors or centroids. In Vector Quantization a large set of feature vectors are taken and a smaller set of measure vectors is produced which represents the centroids of the distribution.

3.1 Speaker Database

The first step is to build a speaker database, $C_{database} = \{C_1, C_2, \ldots, C_N\}$ consisting of N codebooks, one for each speaker in the database. This is done by first converting the raw input signal into a sequence of feature vectors $X= \{x_1, \ldots, x_T\}$. These feature vectors are clustered into a set of M codewords, $C = \{c_1, \ldots, c_M\}$.

The set of codewords is called a codebook. The clustering is done by a clustering algorithm, and K-means clustering algorithm is used for this purpose.

3.2 K-means

The *K*-means algorithm is widely used in speech processing as a dynamic clustering approach. "*K*" is pre-selected and simply refers to the number of desired clusters. One way to compute the code vectors of the training set is to start with an arbitrary random initial estimate of the code vectors and to apply the nearest neighbour condition and the centroid condition iteratively, until a termination criterion is satisfied. The K-means algorithm partitions the *X* feature vectors into *M* centroids. The algorithm first chooses *M* cluster centroids among the *X* feature vectors [7]. Then each feature vector is assigned to the nearest centroid, and the new centroids are calculated. This procedure is continued until a stopping criterion is met, that is the mean square error between the feature vectors and the cluster-centroids is below a certain threshold or there is no more change in the cluster-centre assignment.

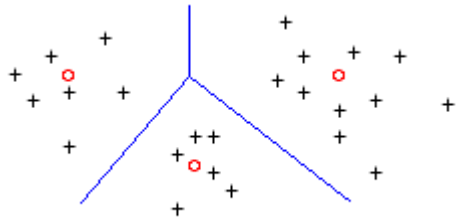

Figure 4: K-means with 3 clusters

3.3 Speaker Matching

In the recognition phase an unknown speaker, represented by a sequence of feature vectors $\{x_1, \ldots, x_T\}$, is compared with the codebooks in the database. For each codebook a distortion measure is computed, and the speaker with the lowest distortion is chosen. One way to define the distortion measure is to use the average of the Euclidean Distances [11]. The Euclidean distance is the ordinary distance between the two points that one would measure with a ruler, which can be proven by repeated application of the Pythagorean Theorem. Thus, each feature vector in the sequence *X* is compared with all the codebooks, and the codebook with the minimized average distance is chosen to be the best.

4. CONCLUSION

The goal of this paper was to discuss a speaker recognition system that could be applied to a speech of an unknown speaker. By investigating the extracted features of the unknown speech and then comparing them to the stored extracted features for each different speaker in order to identify the unknown speaker.

The feature extraction was done by using MFCC (Mel Frequency Cepstral Coefficients). The speaker was modeled using Vector Quantization (VQ). A VQ codebook is generated by clustering the training feature vectors of each speaker and then stored in the speaker database. In this method, the K-means algorithm was used for clustering purpose. In the recognition stage, a distortion measure which based on the minimizing the Euclidean distance was used when matching an unknown speaker with the speaker database. VQ based clustering approach is best as it provides us with the faster speaker identification process.

5. REFRENCES

[1] E. Karpov, "Real-Time Speaker Identification", Master's thesis, University of Joensuu Department of Computer Science, 2003.

[2] J. R. Deller, J. G. Proakis and J. H. L. Hansen, "Discrete-time Processing of Speech Signals", Prentice Hall, New Jersey, 1993.

[3] Jain, A., and Zongker, D., "Feature selection: evaluation, application, and small sample performance". IEEE Trans. on Pattern Analysis and Machine Intelligence 19(1997), 153–158.

[4] K. Sayood, "Introduction to Data Compression", Second Edition, Morgan Kaufmann Publishers, San Francisco, California, 2000.

[5] Mike N., Wei W. (2004), "Speaker Recognition", http://cslu.cse.ogi.edu/HLTsurvey/ch1node47.html

[6] Picone, J. (1993), "Signal modeling techniques in Speech Recognition", IEEE ASSP Magazine, Vol. 81, Issue 9, pp. 1215 – 1247.

[7] Soong, F., A.E., A. R., Juang, B.-H., and Rabiner, L., "A vector quantization approach to speaker recognition". AT & T Technical Journal 66 (1987), 14–26.

[8] T. Kinnunen and P. Franti, "Speaker Discriminative Weighting Method for VQ-Based Speaker Identification", Proc. 3rd International Conference on audio and video-based biometric person authentication (AVBPA), Halmstad, Sweden, 2001.

[9]T. Kinnunen and P. Franti, "Spectral Features for Automatic Text-Independent Speaker Recognition" Licentiate's Thesis. http://www.cs.joensuu.fi/pages/pums/public_results/2004_PhLic_Kinnunen_Tomi.pdf.

[10]http://www.lsv.univsarland.de/dsp_ss05_chap9.pdf

[11]http://en.wikipedia.org/wiki/Euclidean_distance

[12]http://en.wikipedia.org/wiki/Mel_frequency_cepstral_coefficient