Chapter 2

Speaker Diarization and Automatic Language Identification: A Review

This chapter provides an overview of some of the existing methods for speaker diarization and also the methods for automatic language identification. The chapter is organized as follows: Section 2.2 explains the commonly used features for speaker diarization and language identification. Section 2.3 presents a review of modeling techniques for speaker diarization and language identification. The dissimilarity measures used for speech analysis are discussed in Section 2.4. A review of speaker diarization techniques is given in Section 2.5. Section 2.6 presents a review of methods for automatic language identification.

2.1 Introduction

Speaker diarization is the process of automatically partitioning a conversation involving multiple speakers into homogeneous segments and grouping together all the segments that correspond to the same speaker. The first part of the process is known as speaker segmentation or speaker change detection while the second one is called speaker clustering. Hence, speaker change detection followed by speaker clustering is known as speaker diarization.

The automatic language identification (LID) is a process by which the language spoken in a particular speech utterance is identified. Humans are the best LID systems
in the world today. Just by hearing one or two seconds of speech of a familiar language, they can easily identify the language. The performance of any LID system depends on the amount of information and the reliability of information extracted from the speech signal and how efficiently it is incorporated into the system.

2.2 Parameterization of Speech Signals

Speech signals need to be parameterized prior to further processing. Parameterization consists of the extraction of a set of features from the speech waveform, which may present two main characteristics: they must provide a reasonable and compact representation of the speech signal and they must have adequate discrimination capabilities for distinguishing between sounds. Most commonly used features for speaker diarization are mel frequency cepstral coefficients (MFCC), linear prediction coefficients, cepstral coefficients and linear prediction residual signals. For language identification, MFCC, shifted delta cepstral features, pitch and duration are commonly used.

2.2.1 Mel Frequency Cepstral Coefficients

The discrete Fourier transform (DFT) based cepstral coefficients are computed by computing inverse discrete Fourier transform (IDFT) of the log magnitude short-time spectrum of the speech signal. The mel-warped cepstrum is obtained by inserting an intermediate step of transforming the frequency scale to place less emphasis on higher frequencies before computing the IDFT. The mel scale is based on human perception of frequency of sounds [27]. Most of the current speaker diarization systems and language identification systems use mel frequency cepstral coefficients to represent the speaker and language information present in the speech signal.
### 2.2.2 Linear Prediction Coefficients

The theory of linear prediction (LP) is closely linked to modeling of the vocal tract system, and relies upon the fact that a particular speech sample may be predicted by a linear weighted sum of the previous samples [28]. The number of previous samples used for prediction is known as the order of prediction. The weights applied to each of the previous speech samples are known as linear prediction coefficients (LPC). They are calculated so as to minimize the prediction error.

### 2.2.3 Cepstral Coefficients

In many applications, Euclidean distance is used as a measure of similarity or dissimilarity between feature vectors. The sharp peaks of the LP spectrum may produce large errors in a similarity test, even for a slight shift in the position of the peaks. Hence, linear prediction coefficients are converted into cepstral coefficients using a recursive relation [29]. Cepstral coefficients represent the log magnitude spectrum, and the first few coefficients model the smooth envelope of the log spectrum [27]. These coefficients can be obtained either from linear prediction coefficients or from the IDFT of log magnitude spectrum of the speech signal. In both cases, the process results in estimating the vocal tract system characteristics from the speech signal [28].

### 2.2.4 Linear Prediction Residual

The individuality of a speaker associated with the excitation of the vocal tract has been a subject of interest in speaker segmentation studies. Linear prediction analysis models the parameters of the vocal tract system, and hence, the information about the excitation source of the vocal tract is present in the residual signal. In [30], it has been shown that the residual signal contains significant speaker specific knowledge. In [31], an approach for extracting speaker-specific information present in the short segments
of the LP residual is described.

2.2.5 Shifted Delta Cepstrum

The shifted delta cepstral (SDC) features have been introduced to improve the LID performance with respect to the classical cepstral and delta cepstral features [32]. The SDC coefficients are computed, for a cepstral frame at time $t$, according to:

$$\Delta c_n(t, i) = c_n(t + iP + D) - c_n(t + iP - D), \quad n = 0, \ldots, N_f - 1, \quad i = 0, \ldots, k_b - 1 \quad (2.1)$$

where $n$ is the $n^{th}$ cepstral coefficients, $D$ is the advance and delay for the delta computation, $P$ is the time shift between consecutive blocks, and $i$ is the SDC block number. The final feature vector is obtained by concatenation of $k_b$ blocks of $N_f$ parameters.

2.2.6 Pitch

Pitch information also contributes to the uniqueness of the speaker’s voice at the segmental (10 to 30 msec) and suprasegmental (>100 msec) levels. Pitch frequency is the acoustic correlate of the rate of vibration of the vocal cords. The uniqueness of the rate of vibration of the vocal cords is due to the differences in the size of the vocal cords among the speakers, and also due to the speaking style or the accent imposed by the speaker. The physiological constraints determine the average pitch of the speaker over the entire utterance. In general the average pitch of female speakers will be higher than those of male speakers. This is due to the fact that the female vocal cords are much thinner in comparison with those of male speakers.

2.2.7 Duration

As one of the prosodic features, it is believed that the phoneme duration statistics provide language discriminative information.
2.3 Pattern Recognition Models

2.3.1 Gaussian Mixture Models

The basis for using Gaussian mixture model (GMM) is that the distribution of feature vectors extracted from an individual’s speech data can be modeled by a mixture of Gaussian densities. For a $N_f$ dimensional feature vector $\mathbf{x}$, the mixture density function for class $s$ is defined as

$$p(\mathbf{x}/\lambda^s) = \sum_{i=1}^{M} \alpha^s_i f^s(\mathbf{x})$$

The mixture density function is a weighted linear combination of $M$ component unimodal Gaussian densities $f^s_i(\cdot)$. Each Gaussian density function $f^s_i(\cdot)$ is parameterized by the mean vector $\mu^s_i$ and the covariance matrix $\Sigma^s_i$ using

$$f^s_i(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^{N_f}/|\Sigma^s_i|}} \exp\left(-\frac{1}{2}(\mathbf{x} - \mu^s_i)^T(\Sigma^s_i)^{-1}(\mathbf{x} - \mu^s_i)\right),$$

where $(\Sigma^s_i)^{-1}$ and $|\Sigma^s_i|$ denote the inverse and determinant of the covariance matrix $\Sigma^s_i$, respectively. The mixture weights $(\alpha^s_1, \alpha^s_2, \ldots, \alpha^s_M)$ satisfy the constraint $\sum_{i=1}^{M} \alpha^s_i = 1$. Collectively, the parameters of the class $s$ model $\lambda^s$ are denoted as $\lambda^s = \{\alpha^s_i, \mu^s_i, \Sigma^s_i\}$, $i = 1, 2, \cdots, M$. The number of mixture components is chosen empirically for a given data set. The parameters of GMM are estimated using the iterative expectation-maximization (EM) algorithm [33].

For speaker segmentation, GMMs have been built for a fixed set of predetermined acoustic classes, and then the input data stream is classified through a maximum likelihood selection over window scanning.

In language identification task, a GMM is created for each language. Under GMM assumption, the likelihood of a feature vector is represented by a weighted sum of multi-variant Gaussian density. During the recognition, an unknown speech utterance
is represented by a sequence of feature vectors, and the log-likelihood produced by the model is calculated.

2.3.2 Support Vector Machines

Support vector machine (SVM) [34], [35] is based on the principle of structural risk minimization. Support vector machines can be used for pattern classification and non-linear regression. For linearly separable data, SVM finds a separating hyperplane which separates the data with the largest margin. For linearly inseparable data, it maps the data in the input space into a high dimension space $\mathbf{x} \in \mathbb{R}^I \mapsto \Phi(\mathbf{x}) \in \mathbb{R}^H$ with kernel function $\Phi(\mathbf{x})$, to find the separating hyperplane. SVM was originally developed for two class classification problems. The $N_c$ class classification problem can be solved using $N_c$ SVMs. Each SVM separates a single class from all the remaining classes (one-vs-rest approach) [36], [37].

In [38], an SVM classification based supervised technique was proposed for speaker change detection in which they adopted a bottom up binary tree combining three two class SVM classifiers for content based audio segmentation. SVM based supervised technique was proposed in [39] by labeling the speech data around the speaker change as (+1) class and the between speaker change as (-1) class for training an SVM hyperplane and then classified each window as (+1) or (-1). In [40], an unsupervised algorithm using SVM was proposed for speaker change detection.

For language identification, given a set of data corresponding to $N_l$ languages for training, $N_l$ SVMs are trained. Each SVM was trained to distinguish between all data of a single language and all other languages in the training set. During testing, the
class label $y$ of a language pattern $x$ can be determined using (2.2)

$$y = \begin{cases} n, & \text{if } d_n(x) + t_l > 0 \\ 0, & \text{if } d_n(x) + t_l \leq 0 \end{cases}$$

(2.2)

where $d_n(x) = \max \{d_i(x)\}_{i=1}^{N_l}$, and $d_i(x)$ is the distance from $x$ to the SVM hyperplane corresponding to language $i$. The classification threshold is $t_l$.

### 2.3.3 Artificial Neural Network Models

A feed forward artificial neural network (ANN) has an input layer, output layer, and one or more hidden layers. Each layer consists of processing units, where each unit represents the model of an artificial neuron, and the interconnection between two units has a weight associated with it. ANN models with different topologies perform different pattern recognition tasks [41], [42], [43]. Theoretical analysis of multilayer perceptron using sigmoidal activation function suggests that these networks may not draw separating hyperplanes in the feature space [44]. This conclusion exposes the inadequacy of the neural network models using sigmoidal activation function for classification task.

The other ANN model such as autoassociative networks [30] was also explored for speaker segmentation. Autoassociative neural network (AANN) models are used to capture the higher order relations among the samples of the residual signal. Then sample models are generated for every second of voiced speech from the first few seconds of the conversation. These models are used to detect the speaker change points. The characteristics of AANN are described in Section 4.2.

### 2.3.4 Hidden Markov Models

Hidden Markov model (HMM) is a doubly embedded stochastic process where the underlying stochastic process is not directly observable [45]. HMMs can be used as probabilistic speaker models for speaker diarization and probabilistic language models.
for language identification. A HMM not only models the underlying speech sounds but also the temporal sequencing of the sounds. The characteristics of HMM are described in Section 6.3.1.

2.4 Dissimilarity Measures for Speech Analysis

In speaker diarization, the dissimilarities between the acoustic feature vectors are obtained by measuring the acoustic distance. The conventionally adopted measures for this distance are described in this section.

2.4.1 Generalized Likelihood Ratio

Given two acoustic sequences $x$ and $y$, we test whether they were produced by the same Gaussian model (the same speaker) $\lambda^{xy}$ or by two different models (two different speakers) $\lambda^x$ and $\lambda^y$. This question can be answered using the following generalized likelihood ratio (GLR):

$$R_{GLR}(x; y) = \ln p(x|\lambda^x) + \ln p(y|\lambda^y) - \ln p(xy|\lambda^{xy})$$ (2.3)

A high value of $R_{GLR}$ means that the "two-model hypothesis" is more likely than the "one-model hypothesis". The first two terms of $R_{GLR}$ is the log-likelihood of the "two-model hypothesis" and the last term is the log-likelihood of the "one-model hypothesis". A GLR curve is extracted from fixed size adjacent windows that move along the audio signal. The two windows advance frame by frame. Mono-Gaussian models with diagonal covariance matrices are used to build the GLR curve. The maximum peaks of the curve are the most likely speaker change points. A threshold is then applied on the GLR curve to find the speaker changes.
2.4.2 Kullback-Leibler Distance

Let $g_1$ and $g_2$ denote the two multivariate Gaussian distributions, respectively. The Kullback-Leibler distance (KLD) is computed as

$$KLD(g_1, g_2) = \frac{1}{2} tr \left\{ (\Sigma^{-1}_g) + (\Sigma^{-1}_g) \right\} \left( \mu_{g_1} - \mu_{g_2} \right) \left( \mu_{g_1} - \mu_{g_2} \right)^T + \Sigma_{g_1} \Sigma^{-1}_{g_2} + \Sigma_{g_2} \Sigma^{-1}_{g_1} - 2I$$

where $tr(\cdot)$ denotes the trace of a matrix, $\Sigma$ and $\mu$ are, respectively, the covariance matrix and the mean vector of the Gaussian distributions, and $I$ is the identity matrix.

When the window slides over the conversation at the fixed step size, a time varying KLD curve will be obtained. The potential speaker turning points are detected at the peaks of the KLD curve.

2.4.3 Bayesian Information Criterion

The Bayesian information criterion (BIC) is a maximum likelihood criterion penalized by the model complexity (the number of model parameters), widely used for model selection [46]. It can be used for change detection in a sequence of acoustic features.

If $S = \{s_l\}, l = 1, \ldots, L_w$ is a sequence of feature vectors, then the possibility of a speaker change at $l = L_1, (L_1 < L_w)$ can be examined by testing the two hypotheses:

- $H_0$: the entire sequence $S$ is generated by a single speaker and is thus assumed to be represented by a single multivariate Gaussian process $N_S(\mu_S, \Sigma_S)$.

- $H_1$: the sequence $S_1 = \{s_l\}, l = 1, \ldots, L_1$ belongs to one speaker and the sequence $S_2 = \{s_l\}, l = L_1 + 1, \ldots, L_w$ (where $L_w = (L_1 + L_2)$) belongs to a different speaker, represented by two separate Gaussian processes $N_{S_1}(\mu_{S_1}, \Sigma_{S_1})$ and $N_{S_2}(\mu_{S_2}, \Sigma_{S_2})$, respectively.
Here \((\mu_{S_1}, \Sigma_{S_1}), (\mu_{S_2}, \Sigma_{S_2})\) and \((\mu_S, \Sigma_S)\) are the maximum likelihood estimates of the mean vectors and covariance matrices of the three multivariate Gaussian processes, \(N_{S_1}, N_{S_2}\) and \(N_S\) respectively. The difference in the BIC values of the two hypotheses (\(d_{BIC}\) or \(\Delta BIC\)) is given by

\[
d_{BIC} = R - \tau P_f
\]

where

\[
R = \frac{L_1}{2} \ln |\Sigma_{S_1}| + \frac{L_2}{2} \ln |\Sigma_{S_2}| - \frac{L_w}{2} \ln |\Sigma_S|
\]

is the log-likelihood ratio, \(P_f = \frac{1}{2}(N_f + \frac{1}{2}N_f(N_f + 1)) \times \log L_w\) is the penalty factor, where \(N_f\) is the dimension of the feature factor and \(\tau\) is a constant which acts as a threshold for decision making. A positive value of \(\Delta BIC\) indicates that two multivariate Gaussian models best fit the sequence \(\mathcal{S}\), which in turn means that a speaker change is hypothesized at \(l= L_1\). The threshold parameter \(\tau\) is sensitive to channel and other variabilities in the speech data and requires fine-tuning for the type of data under consideration.

### 2.5 Techniques for Speaker Diarization

The speaker diarization task consists of segmenting a conversation involving multiple speakers into homogeneous parts which contain the voice of only one speaker, and grouping together all the segments that correspond to the same speaker.

In [47], Xavier Anguera et al. proposed a many to one preprocessing approach in which they used delay and sum beamforming techniques to fuse the signals from each of the multiple distant microphones into a single enhanced signal. Normally a many to one post processing approach is used, in which a speaker segmentation system is run on each of the microphone channels separately and then the results are merged. In the preprocessing approach, the time delay of arrival (TDOA) between
each of the multiple distant channels and a reference channel is computed incrementally using a window that steps through the signals from each of the multiple microphones. Using the TDOA information, the channels are first aligned and then summed and the resulting enhanced signal is clustered. It is shown that the techniques perform well when compared to the postprocessing approach.

In [48], Jitendra Ajmera et al. present a new approach towards automatic annotation of meetings. Audio recordings are segmented using two independent sources of information: magnitude spectrum analysis and sound sources localization. They are combined using HMM. It is claimed that the combined approach provides promising results and improves the speaker segmentation.

A step by step metric based approach using one class SVM is proposed by [49]. The kernel based dissimilarity measure using one class SVM is applied to both speaker segmentation and clustering. It is claimed that the technique permits the use of any dimensional heterogeneous acoustic feature vectors while keeping the computational cost reasonable and the method is complemented to standard approaches.

Sylvain Meignier et al. [7] have proposed two approaches for speaker diarization. The first one is a classical two step strategy and the second one is an integrated strategy. The step by step system uses a distance based approach namely GLR on different acoustic class individually and the results are merged at the end. For clustering, the same GLR distance is used and the stop criterion is the estimated number of speakers. The integrated system is based on an evolutive hidden Markov modeling (E-HMM) of the conversation. In this iterative approach, both the segmentation and the speaker models are used at each step and are re-evaluated at the next step. These two systems are fused to achieve the lowest speaker diarization error rate.

Apart from the approach proposed in [47], Xavier Anguera et al. [50], have pro-
posed another approach based on beamforming techniques, to obtain a single enhanced signal and speaker-position information from a number of microphones. It is shown that a 25 percent relative improvement is obtained compared to using a single most centrally located microphone.

In [51], the authors have analyzed the correlation between signals coming from multiple microphones and proposed an improved method for carrying out speaker diarization for meetings with multiple distant microphones. The algorithm makes use of acoustic information and information from the delays between signals from the different sources. They claim that the performance has improved significantly when compared to previous systems.

Vishwa Gupta et al. [52] proposed a diarization method by combining speaker clusters using two different feature parameters to get better performance. A fast acoustic change point detection algorithm is first used such that it segments over the data, followed by an iterative Viterbi re-segmentation to refine the segment boundaries. The BIC agglomerative clustering combines the segments into bigger clusters which is followed by a Viterbi segmentation stage using GMMs. The results are reported for telephone conversations.

2.5.1 Speaker Segmentation

Speaker segmentation followed by speaker clustering is called diarization [8], [7], [53], [54]. Speaker segmentation aims at splitting an audio stream into acoustically homogeneous segments, such that every segment ideally contains only one speaker [8], [54]. Speaker clustering refers to unsupervised classification of speech segments based on speaker voice characteristics [54], [55]. That is to identify all speech segments uttered by the same speaker in an audio recording and assign a unique label to them [54], [56].

Various speaker segmentation algorithms have recently been proposed in the liter-
nature [57], [58], [59], [60], [46], [62], [63], [64], [65], [14], [66], [67], [68], [38], [69],
[39]. These segmentation algorithms can be roughly categorized into the following
categories.

2.5.1.1 Decoder Based Segmentation

The most intuitive and primitive segmentation approach assumes that a speaker change
is likely to occur during a region of silence, enabling it to detect each significant silence
that occurs in the input speech stream. A speaker is assigned to each detected silence.
The silence-detection function is performed by a decoder or directly by measuring and
thresholding the speech energy or zero-crossing rate [70], [71], [11], [72], [38]. Notably,
no direct connection exists between a detected silence and an actual speaker change.

2.5.1.2 Model Based Segmentation

The identification of speakers can also be utilized for speaker segmentation. The
model-based segmentation approach assumes that a speaker change is likely to occur
at the time indexes where the model's identification decision change from one speaker
to another. A set of models is derived and trained for different speaker classes such
as female and male speakers, silence, music, laughter, breathing, and noise from a
training corpus. The incoming speech stream is classified using these models. As a
result, prior knowledge is a prerequisite to initialize the speaker models.

Starting from the less complicated case, a universal background model (UBM) is
trained off-line to create a generic speaker model [73], [74], [75]. During segmentation,
this model discriminates between speech and non-speech segments. Since models have
been pre-calculated, the algorithm can be used in real time. Another generic model, the
sample speaker model (SSM), is a predetermined generic speaker-independent model
that is built by sampling the input audio stream [76]. A more complicated case is
the anchor model, where a speaker utterance is projected onto a space of reference speakers [77]. Finally models can be created by means of HMMs [7], [78], [79], [15] or SVMs [80], [81].

Although these segmentation systems have yielded some success, they have some drawbacks. First, the fundamental difficulty in model based approaches is the mismatch in the acoustic conditions of the material adopted for training the statistical models and the data stream being analyzed. Second, the model-based methods require pre-training and therefore the acoustic classes must be defined before the test data are available. However, the acoustic classes in live or recorded streams of speaker-unknown acoustic vocal data, such as real time conversations, meetings and live news broadcasts are constantly changing, making speaker acoustic data for training speaker models difficult to obtain beforehand.

2.5.1.3 Metric Based Segmentation

Segmentation has also been addressed recently using metric based approaches in which two adjacent windows are selected from the speech stream, and their dissimilarities are evaluated. The dissimilarities between the acoustic feature vectors are obtained by measuring the acoustic distance. No knowledge of speakers and training stage is needed a priori. The conventionally adopted measures for this are KLD [82], [11], [83], [84], [85], GLR [57], [59], [71], [65], [68], [86], Mahalanobis distance [82], Bhattacharya distance [82], [85] and BIC [87], [66], [72], [88], [89].

Metric-based approaches based on window scanning have recently become the dominant line of development because of their robustness and effectiveness for speaker change detection. These approaches avert broadcasting errors resulting from the conventional window expansion method. Additionally the window scanning method reduces the computational load, and can detect multiple speaker changes simultaneously.
2.5.1.4 Hybrid Algorithms

Hybrid algorithms combine metric and model based techniques. Usually, metric based segmentation is used initially to pre-segment the input audio signal. The resulting segments are used then to create a set of speaker models. Next, model-based re-segmentation yields a more refined segmentation. In [78], HMMs are combined with BIC. In [72], after having performed an initial BIC segmentation, the acoustic changes that are not found by BIC are detected in a top-down manner, i.e. through a divide and conquer technique.

Another interesting hybrid system is introduced in [7], where two systems are combined namely Laboratoire Informatique d'Avignon (LIA) system, which is based on HMMs, and the Communication Langagiere et Interaction Personne-Systeme (CLIPS) system, which performs BIC-based speaker segmentation followed by hierarchical clustering. The aforementioned systems are with two strategies. The first strategy, called hybridization, feeds the results of CLIPs system into LIA system, whereas the second strategy, named merging, merges preliminary results from LIA and CLIPs system and re-segmentation is performed using the LIA system.

2.5.2 Speaker Clustering

Speaker clustering refers to unsupervised classification of speech segments based on speaker voice characteristics [55]. It is employed to identify all speech segments uttered by the same speaker in an audio recording and assign a unique label to them [56]. Many speaker clustering methods have been developed, ranging from hierarchical ones, such as the bottom-up (also known as agglomerative) methods and the top-down (also known as divisive) ones, to optimization methods, such as the $k$-means algorithm and the self-organizing maps (SOMs) [55], [90]. Speaker clustering approaches are classified
into two main categories: deterministic and probabilistic ones. The deterministic approaches cluster together similar audio segments with respect to a metric, whereas the probabilistic approaches use GMMs or HMMs to model the clusters.

2.5.2.1 Deterministic Methods

Deterministic methods cluster together similar audio segments. Several well-known deterministic techniques are described below:

**SOM-based methods:** SOMs are a powerful tool for speaker clustering. An algorithm for speaker clustering based on SOMs is proposed in [91], [92], [93]. The number of speakers $N_s$ is assumed to be known. The data are divided into short segments. Each segment is considered to belong to only one speaker and to be long enough to enable determining speakers identity. A preliminary segmentation of the audio recordings into speech and non-speech segments is applied using thresholding. Non-speech segments are used to train a non-speech SOM.

Initially, speech segments are randomly and equally divided between the $N_s$ models. The speech segments are clustered into $N_s$ speakers by performing competition between the SOMs. Multiple iterations are allowed during training. After each iteration, the data are re-grouped between the models. The training process is applied again to the new partition until the partitions remain unchanged or their difference between two consecutive iterations is less than a threshold value. At the end of the iterative procedure, the system yields $N_s + 1$ models. $N_s$ models are devoted to the speakers and the last model to the non-speech data.

**Hierarchical methods:** Liu and Kubala propose an on-line hierarchical speaker clustering algorithm [56]. Each segment is considered to belong exclusively to one speaker. The closest pairs of audio segments are found by comparing the distances among all the available segments. To calculate the distance between audio segments
and $s_j$, the GLR is used.

A deterministic step-by-step speaker diarization system is developed by Meignier et al., [7]. It is based on speaker segmentation followed by hierarchical clustering. The number of speakers is automatically estimated. The first step of the algorithm is the macro-class acoustic segmentation that divides the audio into four acoustic classes according to different conditions based on gender and wide-/narrow-band detection. Furthermore, silence and non-speech segments are removed.

The system consists of three modules. The first module performs speaker segmentation using a distance metric approach. The created segments are input to the hierarchical clustering module. Initially, a UBM is trained on the available data. Afterwards, segment models are trained using maximum a posteriori (MAP) adaptation of the UBM. GLR distances are then computed between models and the closest segments are merged until $N_s$ segments are left [7]. Clustering is performed individually on each acoustic macro-class and the results are finally merged. The third module applies the penalized BIC criterion in order to estimate the number of speakers [7].

### 2.5.2.2 Probabilistic Methods

Probabilistic methods use GMMs or HMMs to build models that describe the clusters.

**GMM-based methods:** Many approaches based on GMMs have been proposed. Tsai et al. propose a speaker clustering method which is based on the voice characteristic reference space [55]. The reference space aims at representing some generic characteristics of speaker voices derived through training. The speech features are projected onto a reference space so that they are clustered. The projection vectors reflect the relationships between all segments. They are more robust against the interference from non-speaker factors.

Solomonoff et al. also propose a method for clustering speakers based on GMMs [4].
Each speaker is modeled by a GMM. The models are trained using the EM algorithm to refine the weight and the parameters of each component. A clustering method that is based on maximum purity estimation, which aims to maximize the total number of within-cluster segments from the same speakers, is proposed in [94]. Maximum purity estimation is motivated by the fact that although hierarchical clustering guarantees the homogeneity of individual clusters, it is not guaranteed for all clusters. The method employs a genetic algorithm to determine the cluster where each segment should be assigned to.

**HMM-based methods:** HMMs have been widely used in speaker clustering. Ajmera et al. propose a HMM-based speaker clustering algorithm [95]. Each state of the HMM represents a cluster and the probability density function (pdf) of each cluster is modeled by a GMM. The HMM is trained using the EM algorithm. The initialization of the pdfs is done using the $k$-means algorithm. The technique starts with over-clustering the data in order to reduce the probability that different speakers are clustered into one class. Afterwards, the segmentation is performed using the Viterbi algorithm in each cluster. The next step is to reduce the number of clusters by merging. The clusters are merged according to a likelihood ratio distance measure, such as [96].

Ajmera and Wooters propose another robust speaker clustering algorithm [63]. The algorithm automatically performs both speaker segmentation and clustering without any prior knowledge of the speaker identities or the numbers of speakers. The algorithm uses HMMs, agglomerative clustering, and BIC. The algorithm does not require any threshold and, accordingly, separate training data.
2.6 Techniques for Language Identification

Research in automatic spoken language identification started in early 1970’s. The progress in this area of research was slow for almost two decades. With the advent of public-domain multi-lingual corpora for speech [97], many researchers started showing interest in this area and a lot of efforts and progress have been made.

All the existing language identification systems use some amount of language specific information either explicitly or implicitly and they differ only in the amount of the information used in it. The performance and the complexity of the system are proportional to the amount of linguistic information supplied to the system. During training, some systems require only the speech signal and the true identity of the language. In these systems, language models are derived only from the speech data which are supplied during training.

More complicated language identification systems may require segmented and labeled speech signal of all the languages under consideration. Although the performance of the more complicated language identification systems is superior to others; including a new language into such systems is not a trivial task. Therefore, making a compromise between performance and simplicity has become inevitable if the number of languages under consideration is large. All the spoken language identification systems can be broadly classified into two groups, namely, Explicit and Implicit LID systems [98]. A few representative LID systems in both groups are described here.

2.6.1 Explicit LID Systems

The perceptual experiments done by Muthusamy, et al. [99] on language identification task show that knowledge of a particular language definitely helps in identifying excerpts from it. From this, it can be interpreted that even for an automatic system,
If speech recognizers of all the languages to be identified are used as front-ends, the performance in classifying the languages will be better. For developing a speech recognizer for any language, the basic requirement is the segmented and labeled speech corpus. The LID systems that require speech recognizers of one or several languages, in other words, the systems that require a segmented and labeled speech corpus are termed here as Explicit LID systems. Some of the representative explicit LID systems, available in the literature, are summarized and given below:

- 1993, 1994: Use of phone recognizers as front-end for language identification task was introduced by Lamel and Gauvain in [100] and [101]. Phone recognizers for the languages French and English are built and used in parallel. The unknown speech signal from any of these two languages is processed by the two phone recognizers in parallel. The language associated with the model having the highest likelihood is declared as the language of the unknown speech signal.

- 1994: Berkling, et al. [102] have considered a superset of phonemes of three different languages, namely, English, Japanese, and German, and explored the possibility of finding and using only those phones that best discriminate between language pairs.

- 1994: Andersen, et al. [103] have grouped the total inventory of phonemes into a number of groups, one which contains language-independent phones and three language-dependent phone inventories for the three languages under consideration, and tried to classify the languages.

- 1994: Tucker, et al. [104] have utilized a single language phone recognizer to label multi-lingual training speech corpora, which have then been used to train language-dependent phone recognizers for language identification. They have used three techniques for classifying the languages, namely, the acoustic dif-
ference between phonemes of each language, the relative frequencies of the phonemes of each language and the combination of above two sources of information.

• 1994: Hazen and Zue [105] have pursued a single multi-language front-end phone recognizer instead of language-dependent phone recognizer and incorporated the phonetic, acoustic, and prosodic information derived from speech within a probabilistic framework.

• 1995: Kadambe and Hieronymus [106] demonstrate that the performance of an LID system which is based only on acoustic models can be improved by incorporating higher level linguistic knowledge in the form of trigram and lexical matching. This system is also based on the Parallel Phone Recognition (PPR) approach.

• 1996: Zissman [1] has used a single language-dependent phone recognizer to convert the input speech signal to a sequence of phones and used the statistics of the resulting symbol sequences for language identification, which is termed as Phone Recognition followed by Language Modeling (PRLM). It has been further extended using multiple language-dependent phone recognizers in parallel (Parallel-PRLM) and a reasonable improvement in the language identification performance has been achieved.

• 1997: Navratil and Zhulke [107] have also used a single language-independent phone recognizer but have used improved language models instead of standard bigram models.

• 2003: Prasad [108] has developed a language-independent syllable recognizer for a two language task and tried using it as a front-end, in which the syllable
statistics are used to determine language identity.

- 2003: In [109] Ramasubramanian, et al. studied the PPR system in detail for a 6-language task. A study has been made on three different classifiers, namely, the maximum likelihood classifier, Gaussian classifier, and \( k \)-nearest neighbor classifier, each with three different scores, namely, acoustic score, language-model score, and joint acoustic-language model score and concluded that maximum likelihood classifier with acoustic likelihood score gives best LID accuracy.

- 2006: In [110] Liang Wang, et al. investigated two different approaches that use phonotactic information: Parallel Phoneme Recognition followed by Language modeling (PPRLM) and multi-lingual PRLM. In the PPRLM approach, they modified the system by using four different language models with different discounting methods, including the Linear, Absolute, Good Turning and Witten-Bell. It is shown that the modified PPRLM system with the Witten-Bell discounting outperforms and achieves higher accuracy.

- 2007: Haizhou Li et al. [111] proposed an approach based on vector space modeling (VSM). It is assumed that the overall sound characteristics of all spoken languages can be covered by a universal collection of acoustic units, which can be characterized by the acoustic segment models (ASMs). A spoken utterance is then decoded into a sequence of ASM units. Analogous to representing a text document as a term vector, a spoken utterance was converted into a feature vector with its attributes representing the co-occurrence statistics of the acoustic units. As such, a vector space classifier had been built for LID. This VSM approach leads to a discriminative classifier back end, which is demonstrated to give superior performance over likelihood-based-gram language modeling (LM) back end for long utterances.
2.6.2 Implicit LID Systems

The National Institute of Standards and Technology (NIST) language recognition evaluation conducted in the years 1996 [23] and 2003 [112] shows that the approaches based on a bank of parallel-phone recognizers of multiple languages were the best performing systems. But, some problems can be anticipated with these systems when the number of languages to be identified is increased. In [1], it is shown that parallel-PRLM performs better than PRLM, in which a single phone recognizer is used as a front-end. Further, it is shown that reducing the number of channels (phone recognizers) has a strong negative effect in the performance. From this, one can conclude that the performance of the Parallel-PRLM system is proportional to the number of speech recognizers used in parallel in the system. When the number of languages to be recognized is increased, the number of phone recognizers may need to be increased further. But developing a single phone recognizer with reasonable performance itself is not a trivial task. This clearly shows that attention should be given to developing language identification systems which do not require phone recognizers.

The language identification systems which do not require phone recognizers (or rather segmented and labeled speech data) are termed here as Implicit LID systems. In other words, these systems require only the raw speech data along with the true identity of the language spoken. Here, the language models or the language-specific information are derived only from the raw speech data.

The existing implicit LID systems differ mainly in the feature extraction stage, since the type of feature selected for discriminating languages may be different. Some of the representative implicit LID systems in the literature are discussed below:

- 1977: With an assumption that it is possible to identify gross linguistic categories with great accuracy, House and Neuburg [113] grouped all the speech
samples into five broad linguistic categories and proposed the first HMM based language identification system.

• 1982: Cimarusti and Eves [114] showed that a pattern classifier approach can be used for language identification. The experiment has been conducted using 100 dimensional LPC derived feature vector.

• 1986: Foil [115] examined both formant and prosodic feature vectors and found that formant features were generally superior. The formant vector based language identification system used $k$-means clustering.


• 1991: Muthusamy, et al. [117] suggest a segment based approach to language identification, where the speech is first segmented into seven broad phonetic categories using a multi-language neural network based system with the basic idea that the acoustic structure of any language can be estimated by segmenting the speech into broad phonetic categories.

• 1991: Sugiyama [118] has performed vector quantization classification on LPC features. The difference between the usage of one vector quantization (VQ) code book per language versus one common code book has been explored. In the latter case, languages were classified according to their VQ histograms.

• 1993: Zissman [119] has used a Gaussian mixture classifier for language identification based on the observation that different languages have different sounds and sound frequencies.

• 1994: Li [120] has proposed a system which is based on features extracted at syllable level. In this system, the syllable nuclei (vowels) for each speech utterance
are located automatically. Next, feature vectors containing spectral information
are computed for regions near the syllable nuclei. Each of these vectors consists
of spectral sub-vectors computed on neighboring frames of speech data. Rather
than collecting and modeling these vectors over all training speech, Li keeps
separate collections of feature vectors for each training speaker. During testing,
syllable nuclei of the test utterance are located, and feature vector extraction is
performed. Each speaker-dependent set of training feature vectors is compared
to feature vectors of the test utterance, and most similar speaker-dependent set
of training vectors is found.

• 1999: Pellegrino and Andre-Abrecht [121] have proposed an unsupervised ap-
proach to LID, in which each language vowel system is modeled by GMM trained
with automatically detected vowels. Even though GMM for language identifi-
cation is well-experimented in [1], and [122], the difference here is, the language
models are generated using vowels alone.

• 2000: The first attempt has been made on Indian languages by Jyotsna, et al.
[123]. As [118], VQ based classification has been performed but on 17 dimen-
sional MFCC. It is observed that the acoustic realization of the same sound unit
from different Indian languages is different and some sound units (key sounds)
contribute to the performance of the LID system.

• 2002: In [122], GMM is used to tokenize the speech signal, and language models
are derived from the sequence of tokens. Similar to phone recognizers in the
PRLM system, the GMM tokenizer is trained on one language but is used to
decode information for any candidate language. The decoded sequence is then
used to train bigram models. Further, a parallel GMM tokenization system has
been tried, where for each language, a separate GMM tokenizer is used. It is
shown that as the number of GMM tokenizers is increased beyond a level, there is a degradation in the performance.

• 2003: In [2], Sai Jayaram, et al. have used a parallel sub-word unit recognition approach for LID, where the sub-word unit models are trained without using segmented and labeled speech data. In this work, first, the training utterances are segmented into acoustic segments automatically and clustered using $k$-means algorithm. After clustering, HMMs for each class are generated. The rest of the work is similar to the PPR approach discussed in [1] and [109]. It is claimed that the language identification performance for this system is almost the same as that of the system which uses phone recognizers in parallel [109].

• 2006: A novel approach has been proposed by Nagarajan and Murthy [124] for the LID task which uses parallel syllable-like unit recognizers in a framework similar to PPR approach in the literature with one significant difference. The difference is that the sub-word unit models (syllable-like unit models) for each of the languages to be recognized are generated in an unsupervised manner without the use of segmented and labeled speech corpora. It is demonstrated that, using the acoustic likelihoods of the syllable-like units alone, a reasonable language identification accuracy can be achieved. It is also shown that the language identification performance of the system improves considerably when the acoustic likelihoods based system is combined with the unique syllable-like unit models approach.

• 2006: In [125] a novel dynamic model in ergodic topology was proposed. A segment of pitch contour represented by a set of Legendre polynomial coefficients was successful to the pair-wise language identification task. Feature vectors comprising these polynomial coefficients were formerly modeled by a GMM for
each language. However, the static model like GMM does not take advantage of the temporal information across several pitch contours. It is intuitive that the temporal information of prosodic features should be used for capturing the characteristics of a specific language. Hence, the authors propose a dynamic model in ergodic topology so that it can utilize more information as time proceeds. It is claimed that the performance of language identification has significantly improved the identification rate, even for stress-timed and syllable-timed languages.

- 2006: Felicity Allen et al. [126] have investigated the use of the recently proposed modified group delay function (MODGDF) coefficients in combination with traditional magnitude-based features in a GMM based system. Also the application of feature warping to magnitude-based features and the MODGDF has been examined and found that it can offer a significant cumulative improvement. It is also found that the addition of a modified regression-based SDC further improves the system performance when compared to a more standard SDC configuration.

- 2006: In [32], Bo Yin et al. proposed an approach of combining cepstral features and prosodic features in language identification. The combination approach shows a significant improvement on a GMM-UBM based system which utilizes modern SDC and feature warping techniques. It is proved that the prosodic features are very effective in both tonal and non-tonal LID, as they deliver new language-discrimination information in addition to those from widely used cepstral features.

- 2008: In [127], Khe Chai Sim and Haizhou Li investigated the use of acoustic diversification as an alternative acoustic modeling technique for building a
PPRLM language identification system. This method aims at providing a better acoustic diversification using different acoustic models to provide complementary acoustic emphasis. It is claimed that acoustic diversity is as effective as phonetic diversity in characterizing spoken languages in phonotactic approach to LID. Also they have shown that the acoustic and phonetic diversifications provide complementary language cues.

2.7 Summary

In this chapter, we have presented a review of methods for speaker diarization which includes speaker segmentation and speaker clustering. We have also presented a brief review of the existing methods for automatic language identification. The speech features and the modeling techniques commonly used for these tasks are also discussed. The next chapter addresses a new combination of features for speaker segmentation by fusing the vocal tract features with the excitation source features.