Bi-level Speaker Identification System: Impact on Short Utterances

Leila Beltaifa-Zouari¹, Rania Chakroun²

¹,²Advanced Technologies for Medicine and Signals/Sfax University/Tunisia

ABSTRACT

Despite the recent advances in the field of speaker recognition, there are still many problems for which solutions have to be found. In particular, recognizing a speaker’s identity when only a little amount of speech data is available remains a key consideration since many real world applications have access to speech data having limited duration. In this paper, we present a new approach based on additional information detected from the speech signal to improve the task of automatic speaker identification. In doing so, we highlight how the detection of the speaker’s dialect can be explored to address the research problem related to Short Utterance Speaker Recognition (SUSR). Our study suggests that the automatic detection of the dialect of the speaker can be useful to deal a new approach for speaker identification system when training data are limited and test utterances are very short. We perform the proposed approach in two steps: at the beginning the dialect of the speaker is automatically detected then the identity of the speaker is automatically recognized. We achieved remarkable results using TIMIT database and Support Vector Machines technique. In this context, the proposed approach presents an efficient solution for the constraints related to the memory and computational resource limitation in realistic applications, and hence makes possible the use of large datasets containing many speakers.

Keywords: Automatic Dialect Recognition, Short Utterances, Speaker Identification, Support Vector Machines

[1] INTRODUCTION

The main task of speaker recognition is to extract speaker specific properties while minimizing the effect of the other information and recognize the user’s identity based on a given speech sample. Speaker recognition includes speaker verification, speaker identification, speaker segmentation and speaker indexing. In this study, we focus on speaker identification systems which look for determining which voice sample from a set of known voice samples best matches the characteristics of an unknown input voice sample [1][2][4]. Recently, numerous speaker identification algorithms have been developed [3][10]. However, the performances of these systems have usually been drastically degraded when limited data are presented.

The short utterance speaker identification is required to ensure proper access to confidential information, personal transactions, and security related applications. In fact, in a realistic application, there are many circumstances related to the limitation of computing resources, the conditions and the situations in which the speech was obtained that impose the reduction of the amount of speech data. For that, different methods started to develop in order to address the research problems of Short Utterance Speaker Recognition, which is now becoming a major consideration of modern speaker recognition research [21][11].

In this paper, we systematically analyze the effect of duration of speech utterances on speaker identification and we try to decrease the problem of short utterances trough different processes. Indeed, we introduce a new speaker identification architecture profiting from additional information detected from the speech signal which is the dialect of the speaker. The system operates in two steps: first the dialect is automatically detected, then the speaker of the corresponding region is identified.

We show that the regional dialect differences could be a distinctive characteristic that facilitate more the detection of the identity of a person. We propose a new system evaluated with different Support Vector Machines (SVM) kernel
function. We show that further improvements are possible by taking advantage of feature normalization process like Cepstral Mean and Variance Normalization (CMVN) which help to reinforce the speaker characterization when used utterances have a limited duration.

The rest of this paper is organized as follows. In the next section, a short survey of previous work related to short utterance speaker recognition is given. In section 3, the proposed speaker identification approach which depends on dialect information is described. Experimental set-up and results are demonstrated and discussed in section 4 and conclusions are presented in section 5.

[2] PREVIOUS WORK

Short utterance speaker recognition remains a focus of interest of many researches for quite some time. In fact, in a real circumstance, it might be difficult to collect a large amount of data as required by conventional speaker recognition approaches. For example, there might be some conditions which oblige a person to speak only a little amount of speech like his state of health, his character, his situation, etc. In real life, there are many circumstances that permit only to obtain a small amount of clear speech. In fact, speech obtained could be broken, unclear, recorded in noisy situations or contains some breaks and a little amount of real speech.

Recently, the effect of short duration utterance conditions on system performance is re-introduced in the NIST SRE 2012 [14] which led the research community to further concentration on this problem. The attempt of using smaller amount of data leads to great performance degradation when dealing with a speaker recognition system. To overcome this difficulty, a considerable amount of study are going on in order to develop suitable methods when either the given speech is too small or with the aim of using fewer amount of speech to cut computation costs.

Among the earlier works, we found in [11][12][15] that the problem was addressed to the use of i-vector based speaker verification systems when using small speech utterances. More recent works demonstrate that the use of i-vector provides better performance for short duration speaker recognition systems [15][16]. Short duration utterances in both training and test tasks for speaker recognition was considered by a GMM Universal Background Model (UBM) based framework in [1][2]. In their investigation to improve automatic speaker verification, [12] have looked for a way to adapt systems to work with limited amounts of data. They have used GMM and SVM to achieve this goal. Training and testing using 10 seconds of speech utterances on variations of GMM and SVM was presented in [13][25] for speaker verification task. Speaker recognition needs a large amount of speech data, leading to the use of huge files and complicated processing. This has encumbered the speaker recognition technology to be used widely. Researchers have thus leaded to factor analysis [19][21] for speaker verification. As the utterances get shorter, the performance deteriorates. In [19], significant improvements are obtained with factor analysis with segments shorter than 10s for speaker verification purpose. Moreover, factor analysis were explored in [11] for short utterance speaker recognition.

[3] THE PROPOSED APPROACH

Dialect refers to the different ways of pronouncing a language within a community. Furthermore, the dialect is an important element of the identity of people. It expresses their culture and their history. Nowadays, the dialects of the different languages are increasingly being used in the broadcasts, interviews and debates programs. That’s why we intend to develop a new approach taking into account the dialectal variation of the speaker with the aim of using it as supplementary information that helps to facilitate the discrimination between the speakers.

A. System architecture

The architecture of our proposed system is shown in Figure1. Our proposed approach scheme for speaker identification is defined as a succession of two main blocks. The first is a dialect recognition block constructed with acoustic pretreatment module including feature extraction method followed by a recognition module taking advantage of the SVM classifier. This block has the role of detecting the dialect of the speaker which is used later for the second block which concerns speaker identification and include in turn an SVM recognition module able to recognize the identity of the speaker after comparing the characteristics of the test signal with those constructed in the training phase of the system.
The different steps for comparing the input test speech utterance with the known speakers in the database include the following:

1. Parameters extraction and normalization
2. Computing distance to each dialect in the database
3. Dialect identification (least distance)
4. Comparison to the existing speakers in the database
5. Speaker identification

B. Dialect detection

The test speech utterance is fed into the speaker identification system and this generates a set of feature vectors associated with an acoustic pretreatment module; next, a dialect recognition routine is applied to handle to task of dialect speaker detection followed by a speaker identification routine.

B.1) Feature extraction

Many features have been investigated in the literature [14] where Cepstral based features have become the most successful and most popular when applied to speaker recognition applications. These features are Mel-Frequency Cepstral Coefficients (MFCCs). Experiments were performed with Cepstral features extracted with a 25 ms hamming window with 10 ms overlap. 12 MFCC together with log-energy were calculated every 10 ms. These 13-dimensional feature vectors were augmented with delta and double delta coefficients giving 39-dimensional feature vectors. MFCC features were extracted using the Hidden Markov Model ToolKit (HTK) [23].

The variability of the speech signal caused by different factors like speaker identity, gender, transmission channel, utterance length, session or speaking style make this task difficult. That’s why compensation techniques at different levels such as feature or score levels are needed to cope with speech variability. The most successful referred to feature normalization techniques and score normalization methods. In this work, we tried to reduce the effect of the variability of the extracted characteristics from the speech signal from a session to another. Thus, we used the Cepstral Mean Normalization (CMN) to normalize all the Cepstral coefficients. We extend the previous normalization and adopt another kind of normalization which is Cepstral Mean and Variance Normalization (CMVN) to improve speaker discrimination.

B.2) Dialect recognition

The dialect recognition routine comprises two phases which are the learning phase and the testing phase. Since SVM
classifier has proven to be an effective method for speaker recognition, we decide to take advantage of this classifier for our proposed system.

The main design component of an SVM is the kernel. In fact, due to the kernel which is the main design component in SVM, this classifier can find an appropriate metric in the SVM feature space relevant to the classification problem [17]. Although the existence of different kernel functions, the following functions are the most known:

- Linear: \( K(x_i, x_j) = x_i^T x_j \) (1)
- Polynomial: \( K(x_i, x_j) = (\gamma x_i^T x_j + r)^d, \gamma > 0 \) (2)
- Radial Basis Function (RBF): \( K(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2), \gamma > 0 \) (3)

Where \((x_i, x_j)\) are the input data, \( K(x_i, x_j) \) is the kernel function, and \( \gamma, r \) and \( d \) are kernel parameters.

We have modeled the dialect identification problem using the SVM technique. We have shown that different kernel functions can be used for this classifier. These kernels can be optimized by simply modifying the kernel function parameters. For example, for an RBF kernel having \((\gamma, C)\) parameters, we varied the values of these parameters so we can obtain an optimal learning. Below we summarize the algorithm of dialect recognition training:

Algorithm 1: Dialect learning

Input: speech signal belonging to the dialect \( f \)
Output: dialect model \( mod f \)

1. Extract and normalize Cepstral features
2. Select a kernel function \( K(x_i, x_j) \)
3. Choose the appropriate values of kernel parameters \((C, \gamma, \ldots)\).
4. Set as a stop-learning condition (reach the minimum of error or reach max of iteration)
5. Calculate the output of the learning which is the dialect model \( mod f \).
6. If the number of iterations is reached, learning converge, the learning stops; otherwise another values of kernel parameters are selected and we return to 4.

Since the learning phase serves to acquire the characteristics of every dialect of the database from the extracted parameters, the test phase serves to recognize the unknown dialect of the unknown speaker. In fact, a test utterance is input to the system and Cepstral features are extracted from the speech signal. Thereafter, the model of the test utterance is compared to the dialects models learned with the system with the aim of identifying the most suitable dialect.

C. Speaker identification

Since speaker identification intend to distinguish the identity of a person from a user set, so a learning phase is essential to make an appropriate model for each speaker presented in the database. The model of the test speech utterance of the unknown speaker which is input to the system and its appropriate dialect have been identified is compared with the model of the speakers presented in the database. This additional information serves to facilitate the research of the appropriate speaker while the user number is large, which improves the system performance and makes the short utterance sufficient for training and detecting the speaker’s identity. The following algorithm summarizes the speaker identification process after detecting the speaker’s appropriate dialect.

Algorithm 2: Speaker identification

Input: - speech signal belonging to the speaker \( S 
- dialect identity: Identity(D) 
Output: Identity(S) 

Extract cepstral features
Normalize cepstral features
For \( i = 1, \ldots, ND \)
If \( i = \text{Identity}(D) \)
For \( j = 1, \ldots, N_S \)

\[ \text{Identity}(S) \leftarrow \text{selecting the most suitable speaker identity among} \]
\( N_S \) speakers with SVM classifier
End
End

In the description of the algorithm given above which has the aim of recognizing speakers through their appropriate regions, \( S \) represents the unknown speaker, \( \text{Identity}(D) \) is the dialect of the unknown speaker deduced after extraction of the features and then dialect recognition training and testing with all dialect models \((mod f)\), \( ND \) denotes the number of classes of dialects \( D \), and \( N_S \) represents the number of classes of speakers belonging to the dialect \( i \). \( \text{Identity}(S) \) is determined by selecting the most suitable speaker identity among \( N_S \) speakers with SVM.

[4] EXPERIMENTS AND RESULTS

The experiments were carried out using the TIMIT Database. This corpus contains speakers from 8 major dialect regions of the United States. Each speaker pronounces 10 different sentences with a mean duration of 3.28 seconds per utterance. We examine the performance of speaker identification systems by carrying out experimental evaluations as follows: an SVM baseline system was implemented and evaluated under different scenarios then to be able to make a comparison with previous works, we followed the protocol suggested in [3]. All our evaluations were carried out using 64 speakers from all the regions of TIMIT database. The speakers were equally balanced between the 8 dialects (New England, Northern, North Midland, South Midland, Southern, New York City, Western, and Army Bratand). The 64 speakers were selected as 32 male and 32 female speakers with 4 male and 4 female speakers from each dialect region.

A. Baseline system

The 10 utterances of each speaker were divided into 8 utterances (~24 seconds) for training (two SA, three SX and three SI sentences) and the remaining 2 utterances (two SX sentences) which duration is about 6 seconds for the test task.

The experiments were evaluated with two SVM kernel functions which are the RBF kernel and the linear kernel. For each kind of kernel, we varied its parameters to find the appropriate values for an optimal learning. After achieving the learning phase, we made a set of experiences in the phase of test to evaluate the performance of our systems.

We examined the performance of the speaker identification system with 39-dimension MFCC, CMN and CMVN feature vectors. Results obtained from the different set of experiences are listed in Table 1.

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Linear kernel</th>
<th>RBF kernel</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFCC</td>
<td>92.18</td>
<td>100</td>
</tr>
<tr>
<td>CMN</td>
<td>92.18</td>
<td>100</td>
</tr>
<tr>
<td>CMVN</td>
<td>92.18</td>
<td>100</td>
</tr>
</tbody>
</table>

We notice that the use of the baseline SVM system with RBF kernel achieve an identification rate of 100 % with MFCC feature vectors. The use of CMN and CMVN features keep the same system performance as MFCC features.

The RBF kernel gave better results than the linear kernel. These performances can be explained by the fact that this kernel nonlinearly maps samples into a higher dimensional space than the linear kernel [17]. Furthermore, the use of RBF kernel is a reasonable choice for systems looking for reducing the complexity of the system. In fact, it has lower number of hyper parameters which influences the complexity of model selection. In addition, the RBF kernel has fewer numerical difficulties during the calculation [17]. For these reasons, we choose to handle the remaining experiments with the RBF kernel.
B. Short utterance systems

In order to examine the behavior of our speaker identification systems with short utterances, we prepared a set of utterances having shorter training and testing speech duration segments. Experiments were then conducted with 10s, 8s, 6s and even 4s per speaker for the training task and 3s, 2s, 1s and 0.5s per speaker for the test task. Table 2 presents the best results obtained using the CMVN parameters (which are better than the MFCC ones in this case).

### Table 2: Speaker identification rate for different training and testing durations.

<table>
<thead>
<tr>
<th>Training Duration</th>
<th>Test Duration</th>
<th>3s</th>
<th>2s</th>
<th>1s</th>
<th>0.5s</th>
</tr>
</thead>
<tbody>
<tr>
<td>10s</td>
<td></td>
<td>100.0</td>
<td>98.43</td>
<td>89.06</td>
<td>65.62</td>
</tr>
<tr>
<td>8s</td>
<td></td>
<td>98.43</td>
<td>95.31</td>
<td>84.37</td>
<td>56.25</td>
</tr>
<tr>
<td>6s</td>
<td></td>
<td>96.87</td>
<td>92.18</td>
<td>82.81</td>
<td>62.50</td>
</tr>
<tr>
<td>4s</td>
<td></td>
<td>92.18</td>
<td>85.93</td>
<td>79.68</td>
<td>53.12</td>
</tr>
</tbody>
</table>

The values presented in Table 2 are reported in Figure 2 for better representation.

![Figure 2: Speaker identification Rate for different training and testing durations](image)

We notice that when we reduce either the training or the test data the identification rate decreases. The best result achieved with the SVM-based system with 10s of training and 3s of testing is 100% of identification.

The reduction of the amount of training data to 8s decreases the system performance and gives 98.43% of identification with 3s of testing and only 56.25% of identification with 0.5s of testing.

C. Bi-level speaker identification

To investigate the performance of our proposed approach, we started with evaluating the ability of our system to detect the dialect of the speakers.

C.1) Dialect recognition

The performance of the automatic dialect identification system was evaluated with MFCC, CMN, and CMVN feature vectors. The dialect identification rates obtained from the different set of experiences are listed in Table 3.
Table 3: Dialect identification rate with different feature vectors.

<table>
<thead>
<tr>
<th>Features</th>
<th>Dialect identification rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFCC</td>
<td>98.44</td>
</tr>
<tr>
<td>CMN</td>
<td>98.44</td>
</tr>
<tr>
<td>CMVN</td>
<td>100</td>
</tr>
</tbody>
</table>

The use of MFCC feature vectors gives 98.44% of correct dialect identification rate. Results are then improved with the use of CMVN feature vectors and the system achieves 100% of dialect identification rate.

C.2) Speaker identification

We examined the behavior of the proposed speaker identification systems with short speech segments after the speaker dialect recognition. Experimental results obtained from the evaluation of the dialectal approach with MFCC and CMVN features are presented in Table 4 and Table 5. We recall that in the case of MFCC parameters, the dialect identification rate is 98.44% and in the case of CMVN parameters the dialect identification rate is 100%.

Table 4: Speaker identification rates for Dialectal short utterance system using MFCC parameters

<table>
<thead>
<tr>
<th>Training Duration</th>
<th>Test Duration</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>3s</td>
</tr>
<tr>
<td>10s</td>
<td>100</td>
</tr>
<tr>
<td>8s</td>
<td>100</td>
</tr>
<tr>
<td>6s</td>
<td>100</td>
</tr>
<tr>
<td>4s</td>
<td>98.43</td>
</tr>
</tbody>
</table>

Table 5: Speaker identification rate for Dialectal short utterance system using CMVN parameters

<table>
<thead>
<tr>
<th>Training Duration</th>
<th>Test Duration</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>3s</td>
</tr>
<tr>
<td>10s</td>
<td>100</td>
</tr>
<tr>
<td>8s</td>
<td>100</td>
</tr>
<tr>
<td>6s</td>
<td>100</td>
</tr>
<tr>
<td>4s</td>
<td>100</td>
</tr>
</tbody>
</table>

The various experiences realized show that the performance of speaker identification systems degrades as the utterance length diminishes.

The results of Table 4, Table 5, and Table 2 have been reported in the figures 3, 4, 5, and 6 to make comparison between the short utterance systems (described in Table 2) and the dialectal short utterance systems (tables 4 and 5). Each figure is related to a training duration.

The experimental results using MFCC parameters show that the dialectal short utterance system achieves superior recognition rates over the short utterance system for all cases when both the amount of training data and testing data are varied. These results clearly prove that the recognition of the dialect represents important information for improving the performance of speaker identification systems. This new system facilitates more the task of identification of the speakers after detecting their appropriate regions when short utterances are used.
Figure 3: Speaker identification rates for ten seconds of training duration and different testing durations

Figure 4: Speaker identification rates for 8 seconds of training duration and different testing durations

Figure 5: Speaker identification rates for 6 seconds of training duration and different testing durations
When we apply the CMVN features with the dialectal system, we notice a great improvement in the system’s performance. In fact, from the results obtained previously, it can be observed that the performance of the system decreases as the duration of the speech diminished. Despite this reduction, the performance of the proposed Dialectal approach using CMVN is still very powerful and succeeds to achieve 100% of identification when the test utterances have a length of 3s and also 2s. However, this performance degrades when very short utterances are used and so the new approach achieves 92.18% with 4s of training and 0.5s of testing which outperforms the short utterance system for which the identification rate is around 39%.

CONCLUSION

The challenges of providing efficient speaker identification for applications with access to only short duration speech utterances remains a key hurdle to the broad adoption of speaker identification systems which presents, in turn, a specific application of speaker recognition systems. In this paper, we present a study investigating how the new approach performs when utterance lengths are significantly reduced. This method has focused on the formulation of an approach looking for new information able to facilitate the identification of speakers with much reduced speech data. We prove with the new system that the recognition of the speaker’s dialect can be useful for short utterance speaker identification. In fact we don’t need to use a large amount of training dataset as in traditional algorithms. Besides, we don’t require long test utterances to recognize the speaker. Moreover, there is no need to incorporate lengthy and complicated calculations to handle the situations of having small amounts of speech data. This is an interesting advantage especially for realistic applications that need to reduce the computational and time complexity of the system and so the memory size of the system.

REFERENCES


