Combination of clean and contaminated GMM/SVM for far-field text-independent speaker verification

Christian Zieger, Maurizio Omologo

Fondazione Bruno Kessler-irst, via Sommarive 18, 38050 Povo, Trento, Italy
zieger@fbk.eu, omologo@fbk.eu

Abstract

This paper addresses the problem of speaker verification under reverberant conditions, using only the signal acquired by a single distant microphone. The proposed system combines four different subsystems. Two of them are Gaussian Mixture Model (GMM) based and the other two are Support Vector Machine (SVM) based. The subsystems that use the same type of classifier differ in terms of models: one is trained with clean speech and the other is trained with noisy and reverberant speech obtained through the contamination of the clean data with the measured impulse responses of the room. The results show that the proposed system outperforms each single subsystem under matched or mismatched conditions.

Index Terms: speaker verification, reverberation, GMM, SVM.

1. Introduction

Text-independent speaker verification checks the claimed identity of a speaker without any prior knowledge about the spoken text [1, 2].

In this work we deal with the speaker verification problem in a distant-talking scenario, which is being investigated by the DICIT project. DICIT is a European Project that addresses the development of advanced technologies for speech and acoustic processing based on multi-microphone devices for hands-free control of TV by voice. In general, the speaker verification system has to tackle situations in which some people are sitting in front of the TV and only one is authorized to its control; therefore the system must be able to recognize or verify the identity of the user and reject all the utterances spoken by unauthorized people in the room. This task is being accomplished by combining speaker identification and speaker localization techniques. More details can be found in http://dicit.fbk.eu.

The main issues with far-field speaker verification are related to the channel mismatch due to impulse response changes corresponding to different speaker positions. Attenuation and reverberation effects often cause the energy overlap of neighboring phonemes.

The problem of speaker verification or recognition in a reverberant environment has not received much attention yet. Only a limited number of works can be found in the literature: [3] addressed reverberation compensation, feature normalization and multiple channel combination with a GMM based identification system; [4] applied a beamforming enhancement technique to improve a single channel GMM based identification system; [5] proposed to reduce the channel mismatch by training speaker models with artificially reverberated data; [6] used SVM-GMM supervector, pitch and formant frequency histograms, cross-channel adaptation to reduce the channel mismatch.

In this work we propose a system combining four subsystems at score level, which is demonstrated to be more robust under reverberant conditions than each individual subsystem. Although the DICIT scenario foresees the use of a nested microphone array for the signal processing and interpretation, for this moment we focus on the single microphone case for a choice of reduced computational load. The combination of GMM and SVM for text-independent speaker verification has been investigated recently [7, 8]. Usually GMM performs very well under matched conditions, but the performance decreases under mismatched conditions. SVM is characterized by a high discriminative power, as it is able to separate non-linear separable classes thanks to a non-linear mapping of the feature space. With these considerations we have chosen to combine two GMM based subsystems and two SVM based subsystems. The major contribution introduced by this paper is the combination of subsystems differing not only in the classifier type, but also in model type [5]. Clean models, trained with clean speech, and contaminated models, trained with speech artificially reverberated with measured impulse responses are both used.

We will show that the synergy of all these four subsystems allows a considerable performance improvement compared to each single subsystem or other possible combinations.

The paper is organized as follows: section 2 gives a description of the whole system, section 3 comments on the experimental results and finally section 4 draws the conclusions and hints some possible future work.

2. System Description

The proposed system works as follows (see Figure 1): first the utterance is coded into an N-dimensional feature vector sequence of length $T$ at the front-end stage, then four scores are computed by four different subsystems and combined through a weighted sum at the fusion stage to obtain a new score, based on which the system decides to reject or accept the utterance according to a predefined threshold.

The four subsystems are: clean GMM and contaminated GMM (trained with clean and contaminated speech respectively); clean SVM and contaminated SVM (trained with clean and contaminated speech respectively).

Each contaminated subsystem is trained with clean speech artificially reverberated with only one impulse response.

2.1. Motivation for this work

The idea of the combination of clean and contaminated models rises from the comparison between the scores given by the con-
Real impulse responses were measured for a possible typical scenario (see Figure 3). The room measures are 4.90x3.40 meters. The reverberation time is 700 ms. The microphone is positioned about 2 meters away from the speakers. Four possible speaker positions (p1, p2, p3, p4) were simulated, so that we can evaluate the system performance using different combinations of positions for training and testing data. We want to train models with one impulse response and then evaluate the system performance under mismatched conditions using testing data for different positions. Let us notice that these impulse responses do not differ so much to change completely the acoustic characteristics of the signals [10].

In the data contamination procedure clean speech is convolved with the four different impulse responses respectively, then previously recorded room background noise is added imposing a SNR of 25 dB.

2.3. Front-end
The front-end codes speech sampled at 16 kHz into a 30 dimensional feature vector sequence composed of 15 Mel Frequency Cepstral Coefficients (MFCCs) with their first derivatives. The analysis window is 20 ms and the analysis step is 10 ms. Pre-emphasis is applied with a first order FIR filter. The feature warping method proposed in [11] is used as feature normalization to reduce the mismatch due to the effects of different channels and noise.

2.4. GMM based subsystem
In the GMM subsystems a set of sentences that are representative of the population is used to train a UBM (Universal Background Model) through the iterative Expectation Maximization (EM) algorithm, then speaker dependent models are adapted using the Maximum A Posteriori (MAP) algorithm applied only on the mean vectors of the UBM [12]. The GMM is a parametric model and is completely defined when the weights \( w_i \), the mean vector \( \mu_i \), the covariance matrix \( \Sigma_i \) and the number of mixtures \( M \) are defined \( (i = 1, \cdots, M) \). In the following a general GMM will be denoted as \( G = (w_i, \mu_i, \Sigma_i) \).

The likelihood scoring function for feature vector \( x \) and model \( G \) is:

\[
p(x|G) = \sum_{i=1}^{M} w_i p_i(x) \quad (1)
\]

where \( p_i(x) \) is the i-th multidimensional Gaussian probability
density function with mean vectors \( \mu_i \) and covariance matrices \( \Sigma_i \).

The score of a feature vector sequence \( X = \{x_1, \cdots, x_T\} \) for speaker \( k \) is defined as the difference between the scores given by the speaker model and the UBM:

\[
S_{GMM_k}(X) = S_{GMM}(X|G_k) - S_{GMM}(X|G_{UBM})
\]  (2)

where \( S_{GMM}(X|G) \) is the average log-likelihood of \( X \) for model \( G \)

\[
S_{GMM}(X|G) = \frac{1}{T} \sum_{t=1}^{T} \log p(x_t|G)
\]  (3)

2.5. SVM based subsystem

SVM is a two-class classifier technique largely used in recent years for speaker verification [13]. It is based on hyperplane separators (support vectors) chosen in order to maximize the distance between the hyperplane and the closest training vectors (maximum margin). The discriminative function computed on test feature vector \( x \) is:

\[
f(x) = \sum_{j=1}^{L} a_j t_j K(x, x_{svj}) + \rho
\]  (4)

where \( K(\cdot, \cdot) \) is the Kernel function, \( t_j \) is the ideal output (1 or -1 depending upon whether the corresponding support vector belongs to the class of the enrolled users or to the class of the impostors). The support vectors \( x_{svj} \), the support vector number \( L \), the coefficients \( a_j \) and the constant \( \rho \) are obtained through an optimization process during the training phase. In the classification stage the feature vector is classified as either belonging to the enrolled user for positive values of this function or the opposite for negative values.

The score of a feature vector sequence \( X \) for speaker \( k \) is:

\[
S_{SVM_k}(X) = \frac{1}{T} \sum_{t=1}^{T} U(f(x_t))
\]  (5)

where \( U(\cdot) \) is the unit step function, whose value is 0 for negative argument and 1 for positive argument. In order to reduce the computational complexity in the training and testing phase, the training data of each speaker are quantized using the K-means clustering algorithm with a predefined number of centroids. Although this operation reduces slightly the classification performance, the trained SVM is characterized by a smaller number of support vectors.

2.6. Score fusion

For score fusion we use Matcher Weighting (MW). Each subsystem score is weighted by a factor inversely proportional to the corresponding Equal Error Rate (EER), so that the weights for more accurate subsystems are higher than those of less accurate subsystems:

\[
u_t = \frac{1}{\text{EER}_t} / \sum_{t=1}^{4} 1/\text{EER}_t
\]  (6)

where EER\(_t\) is the EER for the \( t \)-th subsystem. The fused score is given by the weighted sum:

\[
S(X) = \sum_{t=1}^{4} u_t s_t(X)
\]  (7)

Before fusion the scores are normalized in order to have comparable values. We use the Min-Max normalization: this method maps the raw scores to the [0,1] range.

3. Experimental results

The Apaschi database\(^1\) was used to test our proposed system. The database includes Italian phonetically rich sentences. There are 20 utterances for each speaker. The average utterance length is about 4.8 seconds. 83 speakers were selected for the training of the background model and 40 non-overlapping speakers were chosen for individual model training and testing purposes. 10 utterances for each speaker were selected for the training of speaker dependent models. For each speaker model the remaining 10 utterances are used to test the real user and 10 utterances of each of the other speakers are used to test impostor trials. In total, our test set is composed of 400 enrolled user trials and 15600 impostor trials.

For the clean and contaminated GMM based subsystems the chosen number of mixtures is 512. For the clean and contaminated SVM based subsystems the number of centroids is 100. Radial Basis Function (RBF) was used as the Kernel function. The parameter which determines the RBF width is set to 1. In the training phase we take into account the unbalanced data for the speaker and background modeling by weighting the speaker data more than the background data.

We did not perform an exhaustive search of the best subsystem combination considering all possible combinations of the four subsystems. We proceeded combining sequentially the subsystems from the best one to the worst one ensuring that each new subsystem combination provided better results than the previous one. The experimental framework compares the speaker verification performance under both matched and mismatched conditions. The results in terms of average EER are reported in Figures 4 and 5: the white bars represent the results obtained with the clean and contaminated GMM and their combination; the light blue bars are the results for the clean and contaminated SVM and their combination; the black bars are the results for the contaminated GMM/SVM combination and the complete proposed system.

Comparing the clean and contaminated model performance, one can notice that their combination provides always better results than each single model. In fact the combination of the clean model with the contaminated one is able to recover some errors of the contaminated one and at the same time it introduces a limited amount of new errors. We also measured the probability of misses and false alarms for utterances where the contaminated GMM fails or succeeds. Setting the threshold to the value corresponding to the EER condition, the fused system is able to recover 41% of misses and 22% of false alarms. Moreover, only 0.22% of new misses and 0.51% new false alarms are introduced. Finally, it is worth noting that the EER reduction due to the clean and contaminated model combination is more evident under mismatched conditions than under matched conditions, as shown by Figure 5.

Comparing the results of the different classification subsystems one can observe that GMM performs better than SVM. This does not mean that in general GMM classifiers are better than SVM classifiers for a speaker verification system. This experimental result is also due to the fact that SVM are trained with quantized data and an exhaustive optimization of the SVM parameters was not performed. In general the errors of GMM

\(^1\)http://www.elda.org/catalogue/en/speech/50039.html
and SVM classifiers are largely uncorrelated, for this reason the GMM/SVM combination outperforms each single classifier based subsystem. In fact, the combination of the contaminated GMM and SVM subsystems recovers 83% of misses and 33% of false alarms and introduces only 0.68% of new misses and 0.83% of new false alarms.

Finally, one can notice that the whole system performance under mismatched and matched conditions is very similar, being around 1.9% EER in both cases.

4. Conclusions

In this work we have proposed a system, derived from the combination of four subsystems to deal with the problem of speaker verification under reverberant conditions. The whole system uses the synergy of GMM and SVM classifiers as well as different models trained with clean or contaminated speech. It outperforms each single subsystem under matched or mismatched condition. Moreover, the system does not use any multichannel processing to enhance the speech signal.

For future work the authors intend to explore new features more robust against reverberation and also investigate the extension of the proposed single microphone based system to a multichannel based one. Since commands are very often short in time in the scenario of the hands-free control of TV by voice, some work will also be conducted to study the impact of short testing utterances on the performance of our speaker verification system.

5. References


