Channel Compensation for Speaker Recognition using MAP Adapted PLDA and Denoising DNNs

Fred Richardson1, Brian Nemsick2, Douglas Reynolds1

1MIT Lincoln Laboratory, Lexington, MA USA
2University of California Berkeley, Berkeley, CA USA
frichard@ll.mit.edu, brian.nemsick@berkeley.edu, dar@ll.mit.edu

Abstract

Over several decades, speaker recognition performance has steadily improved for applications using telephone speech. A big part of this improvement has been the availability of large quantities of speaker-labeled data from telephone recordings. For new data applications, such as audio from room microphones, we would like to effectively use existing telephone data to build systems with high accuracy while maintaining good performance on existing telephone tasks. In this paper we compare and combine approaches to compensate models parameters and features for this purpose. For model adaptation we explore MAP adaptation of hyper-parameters and for feature compensation we examine the use of denoising DNNs. On a multi-room, multi-microphone speaker recognition experiment we show a reduction of 61% in EER with a combination of these approaches while slightly improving performance on telephone data.

Index Terms: speaker recognition microphone compensation denoising DNN

1. Introduction

The majority of available speech data that is appropriate for training a state-of-the-art speaker recognition system is telephone channel data. This is because large amounts of speaker labeled multi-session data is required to train a speaker recognition system and the publicly available data that meets this requirement is telephone speech [1, 2]. Unfortunately speaker recognition systems trained on telephone data tend not to perform well when applied to speech recorded over other types of channels such as microphone speech. In this work we evaluate two approaches for building speaker recognition systems trained on Switchboard 1 and 2 telephone speech data [1] that can perform well on conversational microphone speech data from the 2009 Mixer 6 collection [3]. These microphone robust systems are developed using Mixer 1 and 2 parallel conversational microphone data [4, 5, 2] collected from 2003 to 2005 using a different set of speakers, rooms and microphones than the Mixer 6 collection.

The first approach we investigate is inspired by a technique developed for domain adaptation at the 2013 JHU SCALE Workshop. The 2013 domain adaptation challenge (DAC13) used data from Switchboard 1 and 2 to train a system which was then evaluated on Mixer telephone speech. The DAC13 evaluation data consisted of the telephone portion of the Mixer 6 collection. Prior Mixer telephone collections were used for adapting the Switchboard telephone system to the Mixer domain. Both supervised and unsupervised techniques were explored for adapting the speaker recognition systems. In this work we investigate the application of one of these techniques - supervised PLDA MAP adaptation [6] - to adapting a telephony speaker recognition system to microphone channel speech.

Recent work has shown that deep neural networks can be very effective for channel compensation in speech recognition algorithms [7, 8, 9]. The type of channel compensation DNNs are used for falls into three basic categories: waveform compensation [10, 8], feature compensation [11, 12, 13, 8, 9] and multi-condition classification [14, 15, 8, 9]. The first two categories are very similar in that they use a DNN regression to reconstruct some possibly intermediate feature representation from a clean channel using some possibly different feature representation of the same data from a noisy channel. For waveform compensation the reconstructed features from the DNN are used to synthesize a new waveform while for feature compensation the output of the DNN may be used directly or may be transformed into a different feature representation. The last category, multi-condition classification, trains a DNN classifier using the same data across a range of different channel conditions but with the same target class labels. While much of this prior work has focused on using synthetic multichannel data by applying reverbation and noise to clean speech data, in this work we will use the Mixer 1 and 2 parallel microphone collection to train a DNN for feature compensation to improve the performance of the Switchboard 1 & 2 speaker recognition system on the Mixer 6 conversational microphone data.

2. Channel Compensation Techniques

2.1. Denoising DNN

A denoising DNN (see Figure 1) is a neural network regression model trained to reconstruct data from a clean target channel given the same data from a different possibly noisy or reverberant channel or from the same channel as the target. The objective function for the denoising DNN is the minimum mean squared error between the output of the DNN and the target channels data. The denoising DNNs output layer uses a linear activation function (instead of the softmax activation function used for a neural network classifier). For this work we use the Mixer 1 and 2 multi-channel data for training the DNN with the telephone channel as the target data. Both the microphone and the target telephone channels are used as input features to the DNN. Several different architectures are investigated but in all cases the hidden layers of the DNN use the same number of nodes and the sigmoid activation function.
2.2. Denoising DNN I-vector System

The denoising DNN described in Section 2.1 has been used to extract features that are beneficial for a range of different speech technologies and applications. The focus of this work is to use features estimated by the denoising DNN as the input to an i-vector system for channel robust speaker recognition. A simplified block diagram of the hybrid i-vector/DNN system is shown in Figure 2. The i-vector system uses a Gaussian mixture model (GMM) which is often referred to as the universal background model (UBM) to extract zeroth and first order statistics from the input feature vector sequence. A super vector created by stacking the first order statics is transformed down to a lower dimensional sub-space using a linear transformation that depends on the zeroth order statistics (see [16] for details). This transformation requires a total variability matrix $\Sigma$ which is estimated from a large set of super-vectors using an EM-algorithm [16] or PPCA [17].

The i-vector is treated as a single low dimensional representation of a waveform that contains both speaker and channel information. With a sufficient number of recorded speakers and sessions it is possible to estimate a full rank within class covariance matrix ($\Sigma_{wc}$) and across class covariance matrix ($\Sigma_{ac}$). However, it should be noted that estimating a full rank covariance matrix ($\Sigma$) is a complex process that requires having at least the same number of speakers as the i-vector sub-space dimension. An effective technique for computing the likelihood ratio that two i-vectors $z_i$ and $z_j$ come from the same speaker ($\mathcal{H}_s$) or from different speakers ($\mathcal{H}_d$) is probabilistic linear discriminant analysis (PLDA). The PLDA likelihood ratio is given by

$$p(z_i, z_j | \mathcal{H}_s) / p(z_i, z_j | \mathcal{H}_d)$$

which can be computed using the “2 covariance model” described in [18] using the hyper-parameters $\Sigma_{wc}$ and $\Sigma_{ac}$. Another important set of parameters for an i-vector system are the mean vector $\mu$ and whitening matrix $W$ which are used to transform the i-vectors to have a unit normal distribution $\mathcal{N}(0, I)$ over a pool of data before applying length normalization [19]. Typically $\mu$ and $W$ are estimated on the same data used to estimate $\Sigma_{wc}$ and $\Sigma_{ac}$.

2.3. Microphone compensation via MAP adapted PLDA

In [6], supervised MAP adaptation of an i-vector system’s $\Sigma_{wc}$ and $\Sigma_{ac}$ PLDA hyper parameters is shown to perform well for the 2013 domain adaptation challenge task (for more details on the DAC13 task see [20]). It is shown that under a certain set of assumptions the MAP estimate of an adapted covariance matrix reduces to a simple linear combination of the source and target domain covariance matrices. This is referred to as the MAP 2-cov model. We use the same approach to adapt the PLDA $\Sigma_{wc}$ and $\Sigma_{ac}$ covariance matrices estimated on the Switchboard 1 and 2 telephone data to the matrices estimated on the Mixer 1 and 2 microphone data using the MAP formula:

$$\Sigma_{adapt} = \lambda \Sigma_{src} + (1 - \lambda) \Sigma_{mic}$$

As in [6], a single parameter $\lambda$ is used to adapt both $\Sigma_{wc}$ and $\Sigma_{ac}$ and the PLDA whitening parameters $\mu$ and $W$ are estimated only on the target (Mixer 1 and 2 microphone) data.

3. Microphone and Telephone Corpora

The Mixer 1 and 2 and Mixer 6 conversational microphone speech collections were used in this work for evaluating microphone channel compensation techniques for speaker recognition. For the Mixer 1 and 2 data there are 239 speakers (123 female and 116 male) with 1035 sessions (averaging 4.3 sessions/speaker). The sessions were recorded over 8 microphones (see Table 1) and a telephone channel in parallel at three different locations: ICSI, ISIP and LDC (see [4, 2, 5] for details).

In order to train a denoising DNN on Mixer 1 and 2 data, a matched filter was used to time align the data from each microphone channels to the telephone channel. Audio files were rejected if the alignment process failed. At the end of the process a total of 873 sessions out of the 1035 available sessions had data for all channels.

The Mixer 6 microphone collection has data from 546 speaker (280 female and 266 male) with 1400 sessions. There are a maximum of 3 sessions per a speaker (the average is 2.5). The sessions were recorded over 14 microphones listed in Table 2 in two office rooms at the LDC (see [21, 3] for more details). Six microphones were selected for this work based on their distance from the speaker and appear in bold in Table 2 (microphones 02, 04, 05, 08, 07 and 13). We chose to evaluate target and non-target trials only on the same microphone and same room since all sessions from a given speaker in Mixer 6 were recorded in the same room.

Mixer 6 also includes sessions with varying vocal effort (high, low and normal). During the course of this work we found that the performance of the high vocal effort data was particularly poor on the telephone channel. The performance of our baseline system described in Section 4 on the NIST 2010 Speaker Recognition Evaluation (SRE10) - Mixer 1 and 2 and Mixer 6 is summarized in Table 3. Our initial investigation revealed that at least some of the Mixer 6 data appears to have
Chan | Microphone
---|---
01 | AT3035 (Audio Technica Studio Mic)
02 | MX418S (Shure Gooseneck Mic)
03 | Crown PZM Soundgrabber II
04 | AT Pro45 (Audio Technica Hanging Mic)
05 | Jabra Cellphone Earbud Mic
06 | Motorola Cellphone Earbud
07 | Olympus Pearlorder
08 | Radio Shack Computer Desktop Mic

Table 1: Mixer 1 and 2 microphones

Chan | Microphone | Distance (inches)
---|---|---
02 | Subject Lavalier | 8
04 | Podium Mic | 17
10 | R0DE NT6 | 21
05 | PZM Mic | 22
06 | AT3035 Studio Mic | 22
08 | Panasonic Camcorder | 28
11 | Samson C01U | 28
14 | Lightspeed Headset On | 34
07 | AT Pro45 Hanging Mic | 62
01 | Interviewer Lavalier | 77
03 | Interviewer Headmic | 77
12 | AT815b Shotgun Mic | 84
13 | Acoust Array Imagic | 110
09 | R0DE NT6 | 124

Table 2: Mixer 6 microphones

distortion on the telephone channel. Since the high vocal effort speech does not appear to adversely affect the other microphone channels and there are at most 3 microphone sessions per a speaker in Mixer 6, we chose to retain the high vocal effort data for the purpose of evaluating microphone speaker recognition performance.

A test set was created from the Mixer data for evaluating microphone performance with 1,230 target and 224,897 non-target trials for each of the 6 channels (7,371 target and 1,347,686 non-target trials pooled across all microphones). The telephone portion of SRE10 test set was used for evaluating speaker recognition performance on telephone data. The SRE10 test set consists of 7,094 target and 405,066 non-target trials.

4. Experimental Setup

A denoising DNN was trained using either 20 log Mel filter banks (MFB) coefficients spanning a bandwidth 300 to 3,140 Hz or 40 Mel frequency cepstral coefficients (MFCCs) including 20 derivatives coefficients computed using these MFBs. The input to the DNN consist of the MFBs or MFCCs feature vectors stacked in a 21 frame window with 10 frames before and after the center frame which corresponds to the target feature vector.

The i-vector systems in all cases uses a 2048 component Gaussian mixture model and 600 dimensional i-vector subspace. The GMM, T, W, $\Sigma_{wc}$, $\Sigma_{ac}$ parameters are all estimated using the Switchboard 1 and 2 data sets. As mentioned in Section 2.3, for the MAP adapted PLDA systems $m$, $W$ are estimated using the Mixer 1 and 2 data set and the $\Sigma_{wc}$ and $\Sigma_{ac}$ covariance matrices are MAP adapted using a $\lambda$ of 0.5. The baseline system uses 40 MFCC feature vectors with MV normalization. For our experimental results we report both the equal error rate (EER) and minimum decision cost function (min DCF) for a target prior of 0.01.

5. Experiments

The first set of experiments using the MAP adapted PLDA model sweeps the value of $\lambda$ from zero to one. The results of these experiments on the Mixer 6 data set are shown in Figure 3 for each microphone and Figure 4 pooled across all microphones. The baseline performance is attained with a $\lambda$ of zero. It can be seen that a value of $\lambda$ between zero and one gives improved performance over the baseline system for all microphones and for the pooled performance. The pooled performance also reflects the calibration of the system across microphones.

Performance for the baseline system, the MAP adapted system and a range of different DNN systems is presented in Table 4 (EER) and Table 5 (min DCF). In the tables, “AVG” is the average EER across microphones and “POOL” is the pooled performance for scoring all microphones together. The difference between the AVG and POOL results to some extent reflects the calibration of a given system. The first row in Tables 4 and 5 gives the results for the baseline system and the second row is
the MAP adapted PLDA model using a \( \lambda \) of 0.5. The remaining rows demonstrate the impact of the feature vector type, the normalization used and the architecture for the denoising DNN systems.

The MAP adapted PLDA model improves EER by 25% for the average and 39% for the pooled. However, the average DCF is degraded by 5% while the pooled DCF is improved by 14%. It is not clear why the MAP adapted system has inconsistent performance for min DCF but yields a significant gain in EER.

A series of experiments were conducted with various types of features, norming, and DNN architectures (see Tables 4 and 5) and the general conclusions are (1) MFCCs perform better than than MFB features, (2) feature normalization is critical but the type of normalization (MV or Gauss) is not and (3) DNNs with more nodes and layers can yield large gains in performance. The best performing system over all is the DNN trained with MFCCs with MV normalization using 5 layers with 2048 nodes each which reduces the EER by 39% for the average and 55% for the pooled and the min DCF by 24% for the average and 35% for the pooled. The last row of the tables gives the performance for applying MAP adaptation to the best performing DNN in the previous row. While MAP adapted DNN system does yield another 8% relative improvement for the average EER and 13% relative improvement for the pooled EER, the average min DCF is degraded by 16% and the pooled min DCF is degraded by 9%.

As noted earlier, it is important for the denoising DNNs to improve microphone performance without degrading performance on conversational telephone speech. To assess the performance impact of the denoising DNN on telephony data we evaluated the 2048x5 DNN in Tables 4 and 5 on the SRE10 telephone data set. The results of this experiment are given in Table 6. Note that there is actually a small gain in performance for the denoising DNN on SRE10 (a 10% reduction in EER and 7% reduction in min DCF). Table 6 also gives the performance for the MAP adapted PLDA system on SRE10 which is quite poor: the EER is increased 52% and the min DCF is increased by 20%. This degradation of the MAP adapted PLDA system may be partly due to the mismatched whitening parameters \( \mathbf{m} \) and \( \mathbf{W} \) which are trained only on Mixer 1 and 2 microphone data.

### 6. Conclusions

In this work we have presented two approaches to microphone channel compensation - MAP adapted PLDA and denoising DNNs - both of which provide a means of developing a microphone speaker recognition system that can take advantage of the large amounts of available labeled telephony data. Two disjoint parallel microphone data sets were used for developing and evaluating these technologies: Mixer 1 and 2 and Mixer 6. While both methods are shown to be effective, the denoising DNN leads to larger gains in performance on microphone data for both EER and min DCF without sacrificing performance on telephone data. The denoising DNN trained on Mixer 1 and 2 parallel microphone data yields a speaker recognition front end that appear to be very robust across microphone and telephone channels. In future work we will compare our current approach with others.
Figure 5: Baseline PLDA

Figure 6: MAP adapted PLDA with $\lambda = 0.5$

Figure 7: 2048x5 denoising DNN

Figure 8: 2048x5 denoising DNN with MAP adapted PLDA to using synthetic parallel multi-channel data which can be generated in large quantities at a much lower cost than the parallel recording of speech sessions.
7. References


