Using i-Vector Space Model for Emotion Recognition

Rui Xia, Yang Liu

Computer Science Department, The University of Texas at Dallas, Richardson, TX 75080, USA
{rx,yangl}@hlt.utdallas.edu

Abstract

Using i-vector space features has been shown to be very successful in speaker and language identification. In this paper, we evaluate using the i-vector framework for emotion recognition from speech. Instead of using standard i-vector features, we propose to use concatenated emotion specific i-vector features. For each emotion category, a GMM supervector is generated via adaptation of the neural one from a large corpus. An i-vector feature vector is then obtained using each emotion specific GMM supervector. The concatenation of these emotion dependent i-vector features is used as the feature vector in the SVM model for emotion classification. Our experimental results on acted and spontaneous data sets demonstrate that our proposed method outperforms other systems using either static features or GMM supervector features, and that system combination yields additional gain.

Index Terms: emotion recognition, i-vector

1. Introduction

Emotion recognition aims to automatically classify the emotion category from a speech signal. This task has received increasing interest in recent years. There are shared tasks and evaluations for this, such as Interspeech Emotion challenge in 2011 [1], Audio/Visual Emotion Challenge and Workshop in 2011 [2].

In previous work, different acoustic features, such as MFCC, prosodic and speech quality features, have been investigated for emotion recognition. Recently, superior performance has been shown using static features obtained by applying lots of functionals on a large number of low-level descriptors, such as pitch and energy [3]. These features have been used as a strong baseline in the challenge task mentioned above. For classification, methods based on GMM [4] and HMM [5] to model dynamic features have been used. For static features, SVM classifiers are often used with great performance [3]. There are also various previous studies on the effect of units or segments for emotion recognition [6], as well as incorporating lexicon information for improved performance [7].

Recently, total variability i-vector modeling has demonstrated significant performance improvement on both speaker identification [8] and language identification tasks [9]. Some studies also borrowed channel compensation ideas for the emotion recognition task. In [10], Kockmann et al. used intersession variability to adapt the means of GMM. In [11], speaker states are considered as the channel effect, and eigenchannel matrix is estimated and eigenchannel vector is extracted as the new feature set for emotion recognition.

Given the success of using i-vectors for speaker and language recognition tasks, in this paper, we evaluate using the i-vector framework for emotion recognition from speech. Different from previous work using standard i-vector features, we propose to use emotion specific GMMs and extended i-vectors. First we build emotion specific GMM supervectors, which are adapted from a universal background model (UBM) trained from a large corpus of neutral speech. Second, we extract i-vector feature sets for each utterance given different emotion specific GMMs. Then we concatenate these emotion specific i-vector features, and use them as the features for the back-end SVM classifier for emotion recognition. We evaluate our method on different data sets, including acted speech and spontaneous speech. Our results show that our approach works well, outperforming other systems.

The remainder of the paper is structured as follows. Section 2 briefly describes i-vector systems and our proposed method using concatenated i-vector features for emotion recognition. Section 3 presents our experiments. Section 4 concludes the paper and discusses some future work.

2. Method

In this section, we first briefly describe the i-vector system developed by Dehak et al. [12], and then introduce our proposed method using i-vector features for emotion recognition.

2.1. I-vector

I-vector system is a technique to map the high dimensional GMM supervector space (generated from concatenating all the mean values of GMM ) to low dimensional space called total variability space. The main idea is to adapt the target utterance GMM from UBM using the Eigenvoice adaption method introduced in [13]. The target GMM supervector can be viewed as shifted from the
UBM. Formally, a target GMM supervector $M$ can be written as:

$$M = m + Tw$$  \(1\)

where $m$ represents the UBM supervector, $T$ is a low dimensional rectangular total variability matrix, and $w$ is termed as i-vector. Assume the mixture number in GMM is $F$, the acoustic features have a dimensionality of $D$, then the size of the GMM supervector $m$ is $F \times D$. The total variability matrix $T$ can be learned using EM algorithm \[12\]. The i-vectors $w$ has a size of $L$, which is much smaller than the size of the GMM supervector ($F \times D$). $w$ can be directly used as new features for the utterance for various recognition tasks.

2.2. Emotion Recognition using Extended I-vector Features

The basic idea of our proposed method is to represent each instance using concatenated i-vector feature vectors extracted based on emotion specific GMM supervectors, and then use these in the emotion classifiers. Fig 1 shows the extraction of these i-vectors in our method.

![Figure 1: Extraction of extended i-vector features.](image)

The first step is to train emotion specific GMMs. Since speech data labeled with emotional categories is rather limited and may not be sufficient to train GMMs from scratch, similar to \[14\], we use an adaptation method, rather than directly training GMMs from the emotional utterances in the training set. A UBM is first built using a large neutral based corpus. Then, given all the training data $X_i$ for an emotion class ($i$ indicates the class label), we use MAP adaption to adapt the pre-trained UBM to each emotion category and obtain emotion specific GMM models. Here similar to speaker adaptation, only the means are adapted.

The second step is to generate i-vector features. For an utterance, we also use MAP adaptation to obtain its GMM supervector $M$ from the UBM, then we compare it to each emotion specific GMM supervector $m_i$:

$$M = m_i + T_i y_i$$  \(2\)

where $i$ denotes the class label, $T_i$ represents the total variability matrix with respect to $m_i$, and $y_i$ is the i-vector for each utterance. All the utterances from the training set are used for training total variability matrix $T_i$ by Eigenvoice adaption. The major difference with the tradition i-vector extraction process discussed earlier is that instead of using one UBM to estimate one total variability matrix, we train $K$ (number of emotion category) different total variability matrices with respect to $K$ GMMs (corresponding to emotion categories). Each utterance thus has $K$ i-vector sets. The motivation behind this is that, rather than measuring the shifted information $Tw$ from one UBM, for each utterance we hereby obtain its distance from each emotion specific GMM.

After getting $K$ i-vector sets for each utterance, we concatenate these features to form the new extended i-vector $S$:

$$S = [i_{11}, ..., i_{1L}, ..., i_{K1}, ..., i_{KL}]$$  \(3\)

The size of this extended i-vector feature set is $K \times L$ ($L$ is the number of extracted i-vector dimension). Note that this i-vector feature extraction step does not use the emotion class labels of the utterances. For every instance in either training or testing, all the emotion specific GMMs are used.

We use these extended i-vector features for emotion classification. Many classifiers can be used for this purpose.

3. Experiments

3.1. Data

We evaluate our system on two different data sets: acted and spontaneous data.

- Acted: USC AudioVisual data \[15\]. 123 sentences are read by one actress with 4 different emotions: anger, happiness, sadness and neutral. The content of sentences is neutral based. The distribution of the emotion classes is balanced in this data set.
- Spontaneous: Interactive Emotional Dyadic Motion Capture English database \[16\]. We use the released version of this corpus, which contains audiovisual data of 1 male and 1 female actor during their dyadic interaction. It includes 14 scripts and 14 improvisation scenarios. The annotations for
each emotion category are given by multiple annotators. We use the same four emotion categories as the previous corpus. The class distribution in this released version is: 25% angry, 15% happy, 20% sad and 40% neutral. There are 941 utterances in this data set. Note that this data set is still not the most natural spontaneous speech, since the speakers are actors and the speech is script/scenario based.

3.2. Experiment Setup

For the GMM model in the i-vector based system, we use 39 dimensional MFCC features, including static, first and second derivatives of MFCCs, and normalized energy parameter. The background GMM model is trained using the Switchboard corpus, which contains a large amount of neutral speech. We use 128 mixture Gaussians in our experiments. We also evaluated using more mixture components, but did not find performance gain from that. The iteration number for training the $T$ matrix is 10. The size of i-vector is set up empirically.

We compare our proposed method to three other systems.

- **GMM supervector as features**

  First, a UBM GMM model is built from a general database. Then, GMM of each utterance is adapted from the UBM via MAP adaption (only means are adapted). Finally the GMM supervector is formed by concatenating the mean values of GMMs.

- **Static features (openSMILE [17])**

  Static feature vectors have been successfully applied to the emotion recognition task recently. 1584 features used in the Interspeech 2010 Paralinguistic Challenge [18] are extracted using openSMILE toolkit [17]. Table 1 shows the list of these features. Details about these features can be found in [18].

Table 1: Static features: 38 low-level descriptors and 21 functionals.

<table>
<thead>
<tr>
<th>Descriptors</th>
<th>Functionals</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCM loudness</td>
<td>Position max./min.</td>
</tr>
<tr>
<td>MFCC [0-14]</td>
<td>arith. mean, std. deviation</td>
</tr>
<tr>
<td>log Mel Freq. Band [0-7]</td>
<td>skewness, kurtosis</td>
</tr>
<tr>
<td>LSP Frequency [0-7]</td>
<td>lin. regression coeff. 1/2</td>
</tr>
<tr>
<td>F0</td>
<td>lin. regression error Q/A</td>
</tr>
<tr>
<td>F0 Envelope</td>
<td>quartile 1/2/3</td>
</tr>
<tr>
<td>Voicing Prob.</td>
<td>quartile range 2-1/3-2/3-1</td>
</tr>
<tr>
<td>Jitter local</td>
<td>percentile 1/99</td>
</tr>
<tr>
<td>Jitter consec. frame pairs</td>
<td>percentile range 99-1</td>
</tr>
<tr>
<td>Shimmer local</td>
<td>up-level time 75/90</td>
</tr>
</tbody>
</table>

- **Standard i-vector features**

  Here one transform matrix $T$ is used, resulting in one i-vector feature vector for each utterance. These i-vector features are used for emotion classification.

We use 10-fold cross validation for our experiments. When splitting data, we made sure that each fold contains similar number of sentences and emotion categories. For performance measure, we use the average classification accuracy from the 10-fold cross validation. SVM classifiers with RBF kernels are used in all the systems (ours and the above baseline systems).

3.3. Emotion Recognition Results

Table 2 shows the classification results on the two data sets. We can see that on both data, the static feature based approach shows better results than GMM supervector based system, especially on the acted data. Using traditional i-vector features performs slightly better than static features on acted data, but worse on spontaneous one. Comparing to other systems, our proposed method yields the best performance for both data sets. More importantly, our feature size is much smaller (160 on both data sets), compared to the static or GMM supervector features. We examined the detailed classification confusion matrix results and found that our method has consistently better performance for all the emotion classes.

Table 2: Emotion classification results on acted and spontaneous corpora using different systems/features.

<table>
<thead>
<tr>
<th>System</th>
<th>acted</th>
<th>spontaneous</th>
</tr>
</thead>
<tbody>
<tr>
<td>GMM supervector</td>
<td>78.8%</td>
<td>63.0%</td>
</tr>
<tr>
<td>static features (openSMILE)</td>
<td>88.1%</td>
<td>70.1%</td>
</tr>
<tr>
<td>traditional i-vector</td>
<td>89%</td>
<td>68.5%</td>
</tr>
<tr>
<td>proposed i-vector</td>
<td>91.1%</td>
<td>71.3%</td>
</tr>
<tr>
<td>system combination</td>
<td>92.0%</td>
<td>74.6%</td>
</tr>
</tbody>
</table>

The last row in Table 2 shows the system combination results from our proposed method and the system using static features. We used a decision level combination, that is, the posterior probabilities from the two systems are linearly combined. A weight of 0.5 is used for both systems. We can see that there is additional gain from this combination for both data sets, with more improvement on the spontaneous corpus. This can be partly explained by the more competitive performance using static features, in comparison to our system. We also evaluated combining our system with the one based on GMM supervector, and found no gain from that. There are two reasons for this. First, both systems use similar feature sources (i-vectors come from GMM supervectors). Second, the GMM supervector based system has much worse performance than ours, therefore adding it does not yield...
additional gain over our system performance. These suggest our extended i-vector features and the static features are more complementary features.

Table 3 shows the performance of our method with respect to the i-vector size \( L \) on both corpora. Note that since there are 4 emotion classes, the feature size in our approach is \( 4 \times L \). We can see that the system performance depends on the i-vector size \( L \). The best setup is data specific, but in general, when \( L = 20 \) to 60, the best performance is achieved.

Table 3: Performance when varying i-vector dimension \( L \)

<table>
<thead>
<tr>
<th>Dimension ( L )</th>
<th>acted</th>
<th>spontaneous</th>
</tr>
</thead>
<tbody>
<tr>
<td>L=10, feature size=40</td>
<td>89%</td>
<td>68.3%</td>
</tr>
<tr>
<td>L=20, feature size=80</td>
<td>91%</td>
<td>70.3%</td>
</tr>
<tr>
<td>L=40, feature size=160</td>
<td>91.1%</td>
<td>71.3%</td>
</tr>
<tr>
<td>L=60, feature size=240</td>
<td>90.5%</td>
<td>70.9%</td>
</tr>
</tbody>
</table>

4. Conclusion

In this paper we evaluate the feasibility of using i-vector models for the emotion recognition task. In our method, first we train emotion class specific GMMs by MAP adaptation from UBM pre-trained based on a large amount of neutral speech. These emotion specific GMMs are used for training the total variability matrix. Then, we apply i-vector extractor to obtain i-vector sets given emotion specific GMMs and total variability matrices. We concatenate i-vector sets to form a larger dimensional i-vector for each utterance, which is then used as features in the SVM classifier. For evaluation, we use both acted and spontaneous emotional data. Our system outperformed other methods using GMM supervector, static features, or traditional i-vector features. The results show that our proposed i-vector features are better representations for emotion recognition task. In addition, one important advantage of our method is the low dimensionality of our features.

In the future work, we plan to investigate dimensionality reduction techniques, such as Within Class Covariance Normalization (WCCN) and Linear Discriminant Analysis (LDA), which have been used on i-vector systems for speaker identification and language identification tasks. Furthermore, the data used in our evaluation is rather small (in terms of both the size of the data as well as the number of speakers), we will evaluate our approach on other data sets to test its robustness.

5. Acknowledgment

The authors thank Shri Narayanan and Carlos Busso for sharing the USC data. This work is supported by U.S. Air Force Award FA9550-10-1-0388. Any opinions expressed in this work are those of the authors and do not necessarily reflect the views of the funding agency.

6. References


