Speaker Clustering in Emotion Recognition

Ni Ding\textsuperscript{1}, Julien Epps\textsuperscript{1,2}

\textsuperscript{1}The School of Electrical Engineering and Telecommunications, The University of New South Wales, Sydney NSW 2052, Australia
\textsuperscript{2}ATP Laboratory, National ICT Australia, Sydney NSW 2015, Australia

\texttt{ni.ding@student.unsw.edu.au, j.epps@unsw.edu.au}

Abstract

Speaker variability is a known challenge for emotion recognition, however little work has been done on speaker similarity in terms of its contribution to the performance in the emotion classification task. In this paper, we investigate this topic, and find a clear link between speaker proximity and the recognition accuracy. Motivated by this result, emotion based speaker clustering is proposed as a new strategy for speaker adaptation. It involves using speaker proximity to cluster individual speakers’ emotion models in the training set on a per-emotion basis, and adapting the test speaker’s emotion from the closest cluster. A series of tests were conducted to explore how system performance varies with clustering method, the number of clusters and the amount of adapting data. Results on the LDC Emotion Prosody and FAU Aibo Corpora show that this method outperforms speaker bootstrap, both in terms of relieving computation load and producing higher accuracy.

Index Terms: speaker clustering, emotion recognition, acoustic adaptation

1. Introduction

Emotion is a speaker’s psycho-physiological experience when his/her mental state is excited internally or externally. Human speech, one expression of somatic changes, contains rich information for emotion perception. Speech based automatic emotion recognition aims to enable machines to recognize emotional states from human voice, and can be applied to any field where judgments about such states are required, e.g. surveillance or human-machine interaction.

A general emotion recognition system contains front-end speech feature extraction and back-end emotion modeling. However, non-emotion-specific information in speech features contribute to making the emotion models less distinguishable, therefore degrading the performance. Furthermore, an exploration of the extent of confusion between speaker specific emotion classes in [1] indicates that speaker variability is a more significant issue.

A typical solution is to use some sort of speaker normalization to remove speaker variation. For example, mean-standard deviation normalization [2] and feature warping [3] have both proved beneficial. However, as explained in [4], some genuine emotion specific information may be eliminated during the normalization process.

Alternatively, in the case that the speaker under test is willing to provide labeled emotional speech samples, adaptation could be another choice. Our previous work [5] showed that for the small amount of adaptation data, adapting from the closest speaker’s emotion models (‘bootstrap’) is superior to speaker normalization. The results also implied that a bootstrap from two or more speakers might be more suitable than just one. In such a case, however, the heavy workload associated with finding similar speakers and composing \( n \)-speaker models from the training data [5] quickly becomes evident. Even with 1-speaker bootstrap, time taken by searching the best initial speaker models could be considerable in realistic sized corpora.

This paper starts with an experiment that probes speaker variability more deeply, by asking the more fundamental question of whether there is a dependency of emotion recognition accuracy on the similarity between speakers. Subsequently, the idea of clustering training speakers is proposed as a more efficient alternative to speaker bootstrap, and emotion based speaker clustering methods are formed by applying popular speaker proximity measurements and clustering algorithms.

2. Experimental Method

2.1. Emotion recognition system

In order to make the results comparable, we used the same system settings as in [5], i.e. in the front-end, speech was segmented into 20ms frames with 50% overlap, a Hamming window applied to each frame followed by 12 MFCC extraction, then a voicing detector to discard unvoiced frames, while a GMM (Gaussian mixture model) and MAP (Maximum a posteriori) were used as the modelling and adaptation methods in the back-end.

2.2. Speech databases

2.2.1. LDC

The English LDC Emotional Prosody speech corpus [6] (herein ‘LDC’) consists of speech from 7 professional actors trying to express 15 emotions while reading short phrases consisting of dates and numbers. There are on average 10 utterances per speaker per emotion. Data from five emotions, namely anger, sadness, happiness, boredom and neutral, were used.

2.2.2. FAU Aibo

The German FAU Aibo Emotion Corpus [7] consists of spontaneous emotional speech chunks of 51 German children aged between 10 and 13 from two different schools. It was used in the INTERSPEECH-09 Emotion Challenge [8] for a 5-class classification task with data from 26 children from one school comprising the training set and data from 25 children from the other comprising the test set. However, it should be noted that since part of the testing set was used for adaptation, the results reported in this paper are not directly comparable with INTERSPEECH-09 Emotion Challenge.
3. Accuracy vs. Speaker Similarity

With large speaker variability, speakers may have feature vectors from a particular emotion state loosely scattered in the acoustic space. On the other hand, some of them might aggregate because of certain homogeneity (anecdotally humans perceive emotions to be relatively similarly expressed between some different speakers). If this kind of aggregation exists, there should be a relationship between the similarity of speakers A and B and the emotion recognition accuracy that can be attained when speaker B’s models are used to recognize speaker A’s emotional speech.

We investigated this by training GMMs on a per-speaker, per-emotion-state basis. Denote \( \lambda_i^{(j)} = (\{w_i^{(j)}, \mu_i^{(j)}, \nu_i^{(j)}\})_{m=1}^M \) as the GMM for speaker \( j \) emotion \( i \) and \( X_i = \{x_{i1}, x_{i2}, ..., x_{iT}\} \) as the features of emotion \( i \) collected from the test speaker. Speaker similarity was expressed by the average log likelihood across all feature vectors \( L_{LLav} \) in (1) [9].

\[
L_{LLav}(x_i | \lambda_i^{(j)}) = \frac{1}{T} \log \sum_{m=1}^M w_i^{(j)} \mathcal{N}(x_i | \mu_i^{(j)}, \nu_i^{(j)})
\]  
(1)

We selected the \( \lambda_i^{(j)} \) that gave the \( k \)th largest \( L_{LLav} \) among all training models associated with emotion \( i \), and repeated this process for each emotion state until the \( k \)th closest model set \( \{\lambda_i^{(j)}\}_{i=1}^k \) was formed. We used this model set to recognize the emotion state of the test speaker. The resulting accuracy was calculated and associated with mean \( L_{LLav} \) over all emotions.

The first experiment was conducted on LDC in 7-fold leave-one-speaker-out manner (in each fold, a different speaker was used as the test speaker). Since a considerable amount of data is required from the test speaker to precisely determine the speaker similarity, we used 70% of utterances of each emotion from the testing speaker to select the \( k \)th closest models, and the remainder for testing. Likewise, in a second experiment, 50% of chunks from the testing set in FAU Aibo were used for choosing closest speaker models and the remainder for testing. UAR (unweighted average recall) was reported as the recognition accuracy.

Accuracy-\( L_{LLav} \) pairs were plotted on speaker basis on LDC in Figure 1, and averaged across speakers in the testing set on FAU Aibo in Figure 2 (without averaging the plot is too dense). Since the speaker independent (S-IND) system is a suitable baseline result for comparison, the corresponding S-IND results without normalization are shown in both figures.

![Figure 1: Accuracy vs. speaker similarity (\( L_{LLav} \)) on LDC, all accuracy-\( L_{LLav} \) pairs are plotted with different shapes related to testing speaker's identity (right: S-IND results for individual test speakers, without normalization (baseline))](image)

![Figure 2: Accuracy vs. speaker similarity (\( L_{LLav} \)) on FAU Aibo, each points represents the averaged UAR and \( L_{LLav} \) values of the \( k \)th closest speakers \( k \in \{1, 2, ..., 26\} \).](image)

4. Speaker Clustering for Emotion Recognition

Consider a training set containing all speakers’ training data for emotion \( i \); the task is to assign \( J_i \) training speakers into \( N \) clusters.

4.1. Speaker proximity measurement

If two speakers can be represented by pdfs \( \lambda_i^{(1)} \) and \( \lambda_i^{(2)} \), their dissimilarity can be expressed by the symmetric KL divergence distances

\[
D_{KLD}(\lambda_i^{(1)}, \lambda_i^{(2)}) = E_{\lambda_i^{(1)}} \log \frac{\lambda_i^{(1)}}{\lambda_i^{(2)}} + E_{\lambda_i^{(2)}} \log \frac{\lambda_i^{(2)}}{\lambda_i^{(1)}}
\]  
(2)

where \( E_{\lambda_i^{(1)}}[\cdot] \) and \( E_{\lambda_i^{(2)}}[\cdot] \) denote the expected value under \( \lambda_i^{(1)} \) and \( \lambda_i^{(2)} \). Since for GMM there is no closed form of (2), two approximations can be used instead.

The first is a Euclidean distance between supervectors (SV), which gives an upper bound to (2) [10]. The extraction process is:

(i). Train S-IND model \( \lambda_i \) for emotion \( i \)

(ii). Adapt to speaker \( j \)'s model \( \lambda_j^{(i)} \) by mean-only MAP

(iii). Form a supervector by stacking \( w_m \left( \hat{\mu}_m - \mu_m \right) C_m^{-1/2} \) for each mixture component [11]

(iv). Denote \( L \) as the dimension of feature vectors, then the Euclidean distance between \( \lambda_i^{(1)} \) and \( \lambda_i^{(2)} \) is

\[
D_{EU}(\lambda_i^{(1)}, \lambda_i^{(2)}) = \sum_{m=1}^M \sum_{l=1}^L w_m (\hat{\mu}_m - \mu_m)^2 C_m^{-1/2}
\]  
(3)

The second is cross entropy (CE) distance, used in [12], which utilizes the inter-speaker likelihood ratio. In this paper, the cross entropy distance was calculated using \( L_{LLav} \):

\[
D_{CE}(\lambda_i^{(1)}, \lambda_i^{(2)}) = L_{LLav}(x_i^{(1)}, \lambda_i^{(1)}) - L_{LLav}(x_i^{(1)}, \lambda_i^{(2)})
+ L_{LLav}(x_i^{(2)}, \lambda_i^{(2)}) - L_{LLav}(x_i^{(2)}, \lambda_i^{(1)})
\]  
(4)

In (4), \( x_i^{(1)} \) and \( x_i^{(2)} \) are the training data for \( \lambda_i^{(1)} \) and \( \lambda_i^{(2)} \).
4.2. Clustering algorithm
Agglomerative hierarchical clustering constructs a dendrogram from bottom to top according to inter-cluster distance linkage. Single (SNG), complete (COMP) and average (AVE) linkages take the minimum, maximum and average distances of vectors as the distance between two clusters respectively. Clusters are formed by a horizontal cut in the dendrogram that results in leaves equal to the specified number of clusters \( N \).

K-means is a well-known center-based clustering algorithm. We considered two types of K-means initialisation: (i) randomly choosing vector samples from the dataset, repeating \( k \)-means 10 times and selecting the one with smallest within-cluster distance as the output; (ii) using the result of the LBG algorithm [13].

By combining each clustering algorithm with two speaker proximity measurements, a number of clustering methods were formed, as summarized in Table 1.

<table>
<thead>
<tr>
<th>Method</th>
<th>Speaker proximity</th>
<th>Clustering Algorithm</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVHSNG</td>
<td>( D_{EU} )</td>
<td>Hierarchical, single linkage</td>
</tr>
<tr>
<td>SVHCMP</td>
<td></td>
<td>Hierarchical, complete linkage</td>
</tr>
<tr>
<td>SVHAVE</td>
<td></td>
<td>Hierarchical, average linkage</td>
</tr>
<tr>
<td>SVKSAMP</td>
<td>( k )-means, sample start</td>
<td></td>
</tr>
<tr>
<td>SVKLBG</td>
<td>( k )-means, LBG start</td>
<td></td>
</tr>
<tr>
<td>CEHSNG</td>
<td></td>
<td>Hierarchical, single linkage</td>
</tr>
<tr>
<td>CEHCOMP</td>
<td>( D_{CE} )</td>
<td>Hierarchical, complete linkage</td>
</tr>
<tr>
<td>CEHAVE</td>
<td></td>
<td>Hierarchical, average linkage</td>
</tr>
</tbody>
</table>

4.3. Proposed System
We proposed a 5-class emotion recognition system as shown in Figure 3. \( N \) clustered emotion models were trained after the speaker clustering process. For a given set of adaptation data \( a_i \), the best initial cluster model for adaptation corresponding to emotion \( i \), was found by searching the cluster index \( n_{i}^{(\text{op})} \) by (5).

\[
n_{i}^{(\text{op})} = \arg \max_{n} LL_{\text{arg}}(a_i|A_{1}^{(n)})
\]  

(5)

The chosen model \( A_{1}^{(n_{i}^{(\text{op})})} \) was then further MAP adapted and used for classification.

![Figure 3: 5-class emotion recognition system using emotion based speaker clustering](image)

4.4. Experiments
4.4.1. Accuracy vs. number of clusters
The performance of the speaker clustering system in Figure 3 could be affected by two factors: the chosen clustering method and the number of speaker clusters. In this experiment, we fixed the amount of adaptation data to 3 utterances per emotion (LDC) or 10% of utterances (FAU Aibo) to investigate the change of recognition accuracy when varying the number of speaker clusters \( N \) for each clustering method. The range of \( N \) was from one (equivalent to 1-speaker bootstrap) to the total number of training speakers (equivalent to adapting from S-IND models).

Experiments were first run on LDC in a 7-fold leave-one-speaker-out manner with the test speaker’s utterances split in 3:7 for adaptation and test purposes. The average accuracy across 7 folds was plotted as a function of cluster number in Figure 4. Then, the same experiment was replicated on FAU Aibo with 10-fold cross validation (in each fold, a different 10% of chunks from each speaker in the test set was used for adaptation and the remainder for test). The average UAR across all test speakers was plotted in Figure 5.

![Figure 4: Accuracy after adaptation vs. number of cluster (N) on LDC](image)

![Figure 5: Accuracy after adaptation vs. number of clusters (N) on FAU Aibo](image)

It can be seen that for each clustering method, there exists an optimum number of clusters, in the range 2 to 6, having the best performance, which is much higher than the baseline. For FAU Aibo, this value seems to be much less than the total number of training speakers, which implies a significant reduction in the required storage and workload on best-speaker model searching relative to the bootstrap method. Generally speaking, the results produced by similar clustering methods are quite comparable. Among all these methods, hierarchical single-linkage \( D_{CE} \) clustering (CEHSNG) performed well on both databases. However, the results of hierarchical methods with complete
linkage are not as good as single and average linkage, and two k-means methods didn’t perform well on FAU Aibo. This might be because the chosen method was not good at realizing clusters of a particular shape. Since different clustering algorithms suit different data structure, e.g., k-means works well in finding convex-shaped clusters and complete linkage tends to find a compact number of clusters with equal diameters, the nature of the dataset may play an important role in the performance of clustering.

4.4.2. Accuracy vs. amount of adaptation data

Although the results in 4.4.1 favor speaker clustering, they are not sufficient for us to draw an objective conclusion on speaker cluster as to whether it is superior to speaker bootstrap or the method of adapting from S-IND models, since a good method should be able to quickly improve system performance with the amount of adaptation data increasing.

For this reason, a comparison of the accuracy for different numbers of adapting utterances/chunks was conducted on LDC in leave-one-speaker-out manner with the test speaker’s utterances split in 7:3 for adaptation and test purposes, and on FAU Aibo with the testing speakers’ data split in 1:9 for adaptation and test. The best clustering results for both databases are shown in Figures 6 and 7. Speaker-dependent (S-DEP) accuracy for LDC (the result in [5]) was also plotted in Figure 6 to show the performance convergence to that of the S-DEP system.

![Figure 6: Accuracy vs. amount of adaptation data (LDC)](image)

![Figure 7: Accuracy vs. amount of adaptation data (FAU Aibo)](image)

The results confirm the advantage of speaker clustering over bootstrap, and interestingly show a capability to exceed the performance of the S-DEP system. For FAU Aibo, the UAR drop from 0 (S-IND results) to 1 adaptation chunks might be caused by the imbalance of the dataset and fewer prototypes for the adaptation chunks. It is also noted that in Figure 6 the accuracy of CEHCOMP is less than that of the 1-speaker bootstrap in the case of 1 adaptation utterance, and in Figure 7 the UAR of SVHSNG is lowest in the case of 1 and 2 adaptation chunks. One possible reason could be that the closest cluster’s identity was not precisely determined with such a limited amount of adaptation data, so that the adapted cluster models were not discriminative for the recognizing the test speaker’s emotion. Finally, the results can be interpreted in terms of the expected trends of accuracy change since the number of emotion-labeled adaptation utterances might be limited or even zero in many practical situations.

5. Conclusion

Experiments conducted on two databases in this study have suggested a strong relationship between emotion recognition accuracy and speaker similarity, supporting the idea of utilizing speaker similarity for improving the performance of emotion classification in general. Following the proposal of an emotion recognizer based on speaker clustering, experiments on eight clustering methods for two speech databases show that by specifying an optimum number of clusters, a higher recognition accuracy than speaker bootstrap is obtainable. Surprisingly, after adaptation from the speaker-clustering models, this scheme was found to exceed the performance of the S-DEP system, which we had previously assumed to be an upper bound.

6. Acknowledgement

This research was supported by the Australian Research Council through Discovery Project DP110105240.

7. References


