Emotion Clustering Based on Probabilistic Linear Discriminant Analysis

Mahnoosh Mehrabani, Ozlem Kalinli, Ruxin Chen*

Sony Computer Entertainment America

mahnoosh@interactions.net, {ozlem_kalinli, ruxin_chen}@playstation.sony.com

Abstract

This study proposes an emotion clustering method based on Probabilistic Linear Discriminant Analysis (PLDA). Each emotional utterance is modeled as a GMM mean supervector. Hierarchical clustering is applied to cluster supervectors that represent similar emotions using a likelihood ratio from a PLDA model. The PLDA model can be trained with a different emotional database from the test data, with different emotion categories, speakers, or even languages. The advantage of using a PLDA model is that it identifies emotion dependent subspaces of the GMM mean supervector space. Our proposed emotion clustering based on PLDA likelihood distance improves 5-emotion clustering accuracy by 37.1% absolute compared to a baseline with Euclidean distance when PLDA model is trained with a separate set of speakers from the same database. Even when PLDA model is trained using a different database with a different language, clustering performance is improved by 11.2%.

Index Terms:emotion clustering, PLDA

1. Introduction

Recent growth of human computer interaction has motivated increasing research on emotion recognition, i.e., automatically identifying emotional state of a speaker [1, 2]. In addition to the linguistic message or text of ”what is being said”, emotions are encoded in the speech signal as the manner of speaking. Identifying emotional state of users of a system based on their speech improves response of the system. An important application of emotion recognition is in spoken dialog systems [3, 4].

The main components of speech emotion recognition systems are: extracting effective features from speech that characterize emotions, and classification of emotions based on the extracted features [1]. A significant number of studies have explored various acoustic features for emotion recognition [2, 5]. In this study, our main focus is on the pattern recognition or emotion classification/clustering component. Several classifiers have been applied for emotion recognition including Gaussian Mixture Models (GMM), Hidden Morkov Models (HMM), K Nearest Neighbors (KNN), Decision Trees, Support Vector Machines (SVM), and Neural Networks [5, 6, 7, 8, 4, 9].

One of the challenges in emotion classification is to use relevant emotional speech databases to train statistical models for each emotion, and adjust training data to the test scenario [1]. However, in many real applications the available training emotional databases are different from test speech data in various aspects, including types of emotions, speakers, noise, and even language. A few studies have explored unsupervised clustering of emotions and speaking styles to construct databases for audio books and Text To Speech (TTS) systems [10, 11].

In a study by Mehrabani and Hansen [12], a cluster refining method for speaker clustering was proposed based on Probabilistic Linear Discriminant Analysis (PLDA) [13, 14]. It was shown that PLDA modeling improved the speaker clustering performance under mixed speaking styles by identifying speaker dependent subspaces, which are less sensitive to speaking style. Inspired by that, here we propose to use PLDA to identify emotion or speaking style dependent subspaces instead of speaker dependent subspaces. In addition, the present study is based on applying the PLDA likelihood scores as a distance measure to combine clusters in hierarchial clustering, while in [12], PLDA likelihood scores were used to refine the clusters after the baseline clustering.

In this study, a semi-supervised emotion clustering based on PLDA is proposed. The proposed emotion clustering method does not train emotion models. Instead, a PLDA model is trained using a separate emotional data set. Linear discriminant classifiers have been used for emotion recognition [3]. Compared to Linear Discriminant Analysis (LDA), which is commonly applied to maximize the between-class data separation while minimizing the within-class scatter, probabilistic LDA or PLDA is a generative model that can be used for recognition on previously unseen classes [13]. Hence, in this work, the training data for the PLDA model is not required to have the same emotions, speakers, language, etc. as the test data.

The proposed emotion clustering system is based on clustering of GMM supervectors in a PLDA subspace. Instead of directly clustering the extracted feature vectors, each speech unit or utterance is modeled as a GMM, and represented by the GMM mean supervector, which is generated by stacking all the mean vectors of Gaussian mixtures. GMM mean supervectors have been proven to be effective for speaker and language recognition applications [15, 16]. In addition, joint factor analysis and i-vectors have been used to improve speaker recognition for emotional speech by distinguishing between speaker versus emotion subspaces of the GMM mean supervector space [17]. GMM supervectors have also been applied to emotion recognition [18].

The present study proposes an emotion clustering method that directly applies likelihood sores from a PLDA model in hierarchical clustering of emotional speech that was not present in PLDA training. We tested the proposed emotion clustering method with two scenarios. In the first scenario, training data for the PLDA model and test data are selected from the same emotional database but different speakers and different emotions. The second scenario includes training the PLDA on a database and testing it with another database with different speakers and language. In both cases, it is shown that emotion clustering in the PLDA subspace improves speaker independent emotion clustering performance.

The presented PLDA based emotion clustering method has various applications, especially when there is mismatch be-

[*] Mahnoosh Mehrabani was an intern at Sony Computer Entertainment America R&D when this work was done.

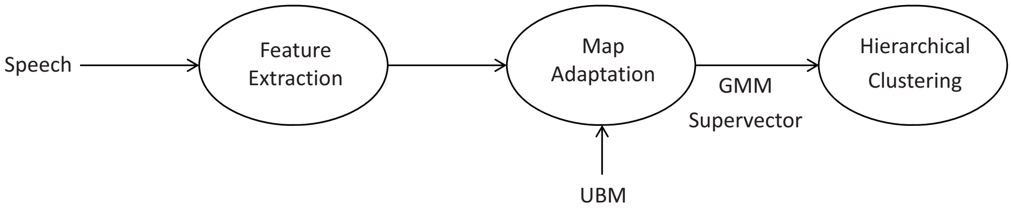

Figure 1: *Flow diagram of baseline emotion clustering.*

tween train and test data sets or when training emotional speech data is not properly labeled. In addition, real emotions are more complex than acted ones and are hardly distinguishable even by humans. However, our proposed system is based on clustering the emotions rather than labeling each utterance with a particular emotion. Therefore, the PLDA model can be trained with acted or exaggerated emotions to learn the supervector dimensions that are more emotion dependent and less speaker or language or text dependent. The trained PLDA model can then be applied to cluster real emotions. Another application of the proposed emotion clustering system is as a front-end to provide homogeneous speech data for multi-style training of other systems such as speech recognition.

Sec. 2 briefly explains the PLDA model that we applied for emotion clustering and Sec. 3 presents the proposed emotion clustering method based on the PLDA model. Sec. 4 describes emotional databases that were used for this study, and clustering results for these databases. Sec. 5 includes the conclusions.

2. Probabilistic Linear Discriminant Analysis

The core idea of the proposed emotion clustering method is identifying emotion related subspaces in the speech data among all the other subspaces that represent variations such as speaker, language, etc. This is done based on Probabilistic Linear Discriminant Analysis (PLDA). Linear Discriminant Analysis (LDA) is a common method in pattern recognition using a linear combination of the features to separate two or more classes based on maximizing the between-class data separation while minimizing the within-class scatter [19]. Probabilistic LDA is a generative model which is more suitable for recognition tasks on previously unseen classes [13]. Therefore, PLDA model can be trained on any available emotional database, and be applied to cluster emotional speech for emotions, speakers, or languages that were not present for training.

The PLDA model that we used for emotion clustering was originally proposed for face recognition applications when the lighting or pose of the probe and gallery images were different [14]:

$$x_{ij} = \mu + Fh_i + Gw_{ij} + \varepsilon_{ij} \tag{1}$$

where x_{ij} represents the $j'th$ image of the $i'th$ individual, with $i = 1,..,I$ and $j = 1,...,J$. This model has two components: the signal component $\mu + Fh_i$ that describes the between-individual variation, and the noise component $Gw_{ij} + \varepsilon_{ij}$ that denotes the within-individual noise. The term μ is the overall mean of the training data, and F and G are matrices, which contain bases for between-class and within-class subspaces, respectively. h_i and w_{ij} are latent variables and finally ε_{ij} is the residual noise term which is defined to be Gaussian with a diagonal covariance matrix Σ [14]. The output of PLDA training is the model $\theta = \{\mu, F, G, \Sigma\}$, which is trained using Expectation Maximization (EM) algorithm.

In this study, each training sample or x_{ij} is a GMM mean supervector, which represents an utterance. $i = 1,..,I$ are the emotions, and $j = 1,...,J$ are various samples for each emotion class, and therefore, matrix F is the matrix that captures the emotion specific subspaces in the supervector space and $\mu + Fh_i$ in Eq. (1) depends only on the emotion or the speaking style of the speaker, while G includes subspaces that contain other information from the speech signal such as the spoken message, speaker, etc. and $Gw_{ij} + \varepsilon_{ij}$ represents all those factors rather than the emotion.

3. Proposed Emotion Clustering

In this study we propose an emotion clustering algorithm, which groups utterances with similar emotions using a previously trained PLDA model. Each utterance is represented by a GMM mean supervector, and clustering is performed in the supervector space. To show the advantages of the PLDA modeling, the results are compared to a baseline system that also performs the clustering on GMM mean supervectors, without using the PLDA model.

Fig. 1 shows the block diagram of baseline emotion clustering. The feature extraction and GMM modeling steps that are depicted in Fig. 1 were also used for the proposed PLDA emotion clustering. First, acoustic features were extracted from every frame of each utterance. Since the focus of this study is on improving pattern recognition algorithms for emotional speech modeling, we used low-level spectral features, MFCCs (13 coefficients including energy, deltas and double deltas), that have been successfully applied for emotion recognition [7, 20]. We acknowledge that using prosodic features can improve the emotion classification performance, however, this study is concentrated on modeling schemes rather than feature selection.

Next, a Maximum A Posteriori (MAP) adaptation was used to adapt a previously trained Universal Background Model (UBM) to the feature vectors extracted from each utterance. Each utterance was represented by a GMM supervector obtained by stacking adapted GMM mean vectors, and clustering was performed in the supervector space.

Fig. 2 shows the block diagram of our proposed emotion clustering system. First, all utterances from the train and test data sets were modeled as GMM supervectors. Next, the training supervectors were used to train a PLDA model. Note that the train data set is not required to have the same emotions or speakers as the test set. The PLDA training data should just include utterances with labeled emotions for a number of different emotions with a variety of samples such as different speakers for each emotion.

Finally, the test supervectors were clustered using hierar-

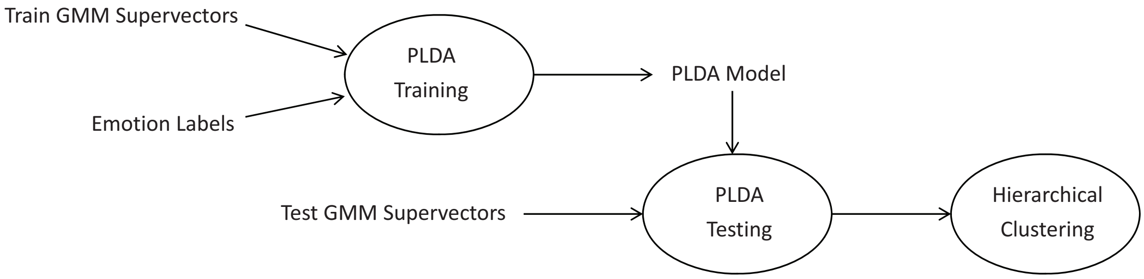

Figure 2: *Flow diagram of proposed PLDA emotion clustering system.*

chical clustering with a likelihood distance based on the trained PLDA model. The hierarchical agglomerative clustering is a bottom-up technique, which starts with each data point being a cluster and subsequently links the closest clusters together until a stopping criterion is satisfied. We applied the PLDA likelihood ratio as the distance measure to compare and link the clusters.

The likelihood ratio that is used as a distance measure for clustering, is the likelihood that two supervectors belong to different emotions divided by the likelihood that two supervectors belong to the same emotion. For two GMM supervectors x_1 and x_2, the PLDA likelihood ratio is:

$$R(x_1, x_2) = \frac{likelihood(diff)}{likelihood(same)} = \frac{P(x_1)P(x_2)}{P(x_1, x_2)} \quad (2)$$

The larger the likelihood ratio, the two supervectors are less likely to be from the same emotion cluster. x_1 and x_2 can be represented based on latent variables using Eq. (1), and if they are from the same emotion class, they have the same latent variable h. The likelihood ratio can be calculated by marginalizing over the hidden variables [13, 14]. Next, results of the proposed PLDA emotion clustering are presented.

4. Results and Discussion

The proposed PLDA emotion clustering was evaluated using three emotional corpora with three different languages. The first database was a Mandarin emotional corpus (THU Emotional) with simulated emotions, but speakers were not actors. THU Emotional corpus contains emotional speech from 30 speakers (15 female and 15 male), and 50 short utterances (approximately 2 sec.) per speaker per emotion. The 5 emotional categories for THU Emotional corpus are: neutral, angry, anxious, happy, and sad. The second emotional database was LDC Emotional Prosody, which is in English. The Emotional Prosody corpus includes 15 emotional categories: disgust, panic, anxiety, hot anger, cold anger, despair, sadness, elation, happy, interest, boredom, shame, pride, contempt, and neutral simulated by 8 professional actors (5 female and 3 male) reading short dates and numbers [21]. The third emotional corpus that was used for this study was Berlin Database of Emotional Speech (EMO-DB) [22]. EMO-DB contains emotional speech from 10 actors (5 female and 5 male) reading 10 German utterances with 7 emotions: anger, boredom, disgust, fear, joy, neutral, and sadness.

We evaluated our emotion clustering method with two scenarios: within and across corpora. For this study we focused on the clustering of 5 emotional categories: neutral, angry, anxious, happy, and sad. For the baseline system, hierarchical clustering was applied directly on GMM supervectors without using the PLDA model (Fig. 1). Euclidean distance and Ward's linkage method [23] were used for hierarchical clustering baseline.

For the first scenario, we evaluated the proposed emotion clustering algorithm on THU Emotional corpus to show the effectiveness of PLDA model to distinguish between speaker dependent and emotion dependent supervector subspaces. The speech data was divided into train and test sets, where train set included speech from 20 speakers (10 female and 10 male), and test set included 10 speakers (5 female and 5 male) with no overlap between train and test speakers. Only the neutral speech from the train set was used for UBM training. All emotional speech from the train speakers with 5 emotion categories was used to train the PLDA model, which will be referred as PLDA-5C. Clustering accuracy was applied to evaluate emotion clustering performance, which is the number of correctly clustered utterances divided by the total number of utterances.

Table 1 compares performance of emotion clustering in PLDA subspace to the baseline using average binary and 5-class clustering accuracies. Average binary clustering accuracy represents mean of clustering accuracy for every pair of emotions. In THU Mandarin emotional database, binary clustering accuracies with the baseline were close to chance for most emotion pairs. As shown in the table, using the PLDA model improved average binary clustering performance by +22.0% absolute. The clustering accuracy for 5 emotions improved by +37.1% absolute with the PLDA model compared to the baseline. The most improved clustering performances were angry-vs-sad and angry-vs-neutral, which achieved 97.3% and 93.7%, respectively. The most challenging pair was angry-vs-happy, where only 51.7% was obtained. Binary clustering accuracies for emotion pairs are compared in Fig. 3 for the baseline and proposed PLDA emotion clustering.

	Baseline	PLDA-5C
Average Binary Clustering Accuracy	54.8%	76.8%
5-Class Clustering	28.0%	65.1%

Table 1: *Average binary clustering and 5-class emotion clustering accuracies for THU Emotional corpus.*

Next, we evaluated the second scenario, where two different databases were used for training and testing. The Emotional Prosody English speech corpus was used as a test set and the EMO-DB German emotional corpus was used to train the

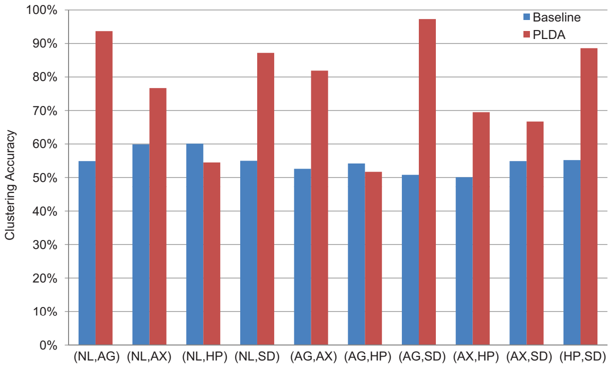

Figure 3: *Binary clustering accuracies for THU Emotional corpus: comparing baseline and proposed PLDA-5C method for every pair of emotions from the set of 5 emotions: neutral(NL), angry(AG), anxious(AX), happy(HP), and sad(SD).*

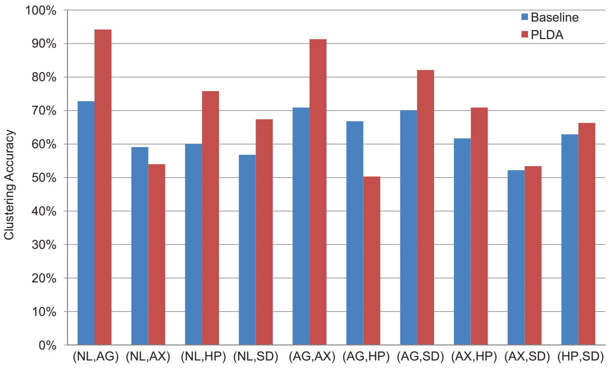

Figure 4: *Binary clustering accuracies for Emotional Prosody corpus: comparing baseline and proposed PLDA-3C method for every pair of emotions from the set of 5 emotions: neutral(NL), angry(AG), anxious(AX), happy(HP), and sad(SD).*

PLDA model. Note that these two corpora had different languages and different speakers. Since EMO-DB did not have enough data to train the UBM, a combination of TIMIT data and EMO-DB was used for UBM training. For this scenario, various combinations of emotions were examined for PLDA model training. The best results were achieved when the PLDA model was trained with three emotional categories of neutral, angry, and sad, which will be referred as PLDA-3C.

	Baseline	PLDA-3C
Average Binary Clustering Accuracy	63.3%	70.6%
5-Class Clustering	34.0%	45.2%
3-Class Clustering	51.2%	70.1%
Neutral-Emotional Clustering	50.7%	60.9%

Table 2: *Baseline and proposed PLDA clustering accuracies for Emotional Prosody corpus (English) when PLDA model is trained with EMO-DB (German).*

Table 2 shows the emotion clustering performance on English emotional prosody database obtained using the baseline method and the proposed PLDA based clustering method, where PLDA model is trained with EMO-DB in German. Compared to the baseline, the proposed PLDA based emotion clustering improves the average binary clustering performance by +7.3% and 5-class emotion clustering performance by +11.2% absolute. The emotional versus neutral speech clustering was also improved by +10.2%. The 3-class emotion clustering for neutral, angry, and sad, which were the emotions used to train PLDA model represent maximum performance improvement of +18.9%. However, binary clustering accuracy for happy and anxious, which were emotions that were not used for PLDA training also increased from 61.7% to 70.9% (+9.2%). This shows the generalization characteristic of the PLDA model, not only for unseen speakers and languages, but also for unseen emotions. For English Emotional prosody database, among pairs of emotions, binary clustering accuracy improved the most for angry-vs-neutral and angry-vs-anxious, achieving 94.2% and 91.3%, respectively. Similar to the THU corpus, the most difficult pair to cluster was angry-vs-happy, where only 50.3% was obtained. Binary clustering accuracies for emotion pairs are compared in Fig. 4 for the baseline and proposed PLDA emotion clustering.

We have also tried clustering English emotional data using a PLDA model trained with THU Mandarin emotional data. Our initial experiments showed that PLDA model trained with Mandarin emotional data did not improve emotional clustering performance for English. This can be since Mandarin and English are not from the same language family and also Mandarin is a tonal language.

5. Conclusions

The task of unsupervised emotion clustering, or grouping speech utterances with similar emotions is a very challenging task. Alternatively, training statistical models for each emotion requires labeled data from several speakers. However, emotional databases generally include acted or exaggerated emotions, and models trained with such data might not be useful to classify real data. In this study, we proposed a semi-supervised approach, which uses a PLDA model for emotion clustering, and showed the significant performance improvement compared to unsupervised clustering. Each utterance was first, represented by a GMM mean supervector, and PLDA model was used to distinguish emotion discriminative subspaces and use them for emotion clustering of unseen emotions, speakers, and even languages.

Future research can include a wider variety of features, emotions, and databases. The proposed clustering method can also be applied as a front-end for speech recognition systems to improve the performance for emotional speech by clustering speech data with similar emotions.

6. References

[1] T. Vogt, E. André, and J. Wagner, "Automatic recognition of emotions from speech: a review of the literature and recommendations for practical realisation," in *Affect and Emotion in Human-Computer Interaction*, 2008, pp. 75–91.

[2] S. G. Koolagudi and K. S. Rao, "Emotion recognition from speech: a review," *International Journal of Speech Technology*, vol. 15, no. 2, pp. 99–117, 2012.

[3] C. M. Lee and S. S. Narayanan, "Toward detecting emotions in spoken dialogs," *Speech and Audio Processing, IEEE Transactions on*, vol. 13, no. 2, pp. 293–303, 2005.

[4] S. M. Yacoub, S. J. Simske, X. Lin, and J. Burns, "Recognition of emotions in interactive voice response systems." in *INTERSPEECH*, 2003.

[5] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition*, vol. 44, no. 3, pp. 572–587, 2011.

[6] B. Schuller, G. Rigoll, and M. Lang, "Hidden markov model-based speech emotion recognition," in *IEEE ICASSP*, vol. 2, 2003, pp. II–1.

[7] B. Vlasenko and A. Wendemuth, "Tuning hidden markov model for speech emotion recognition," in *33th Fortschritte Der Akustik*, 2007, pp. 317–320.

[8] B. Schuller, G. Rigoll, and M. Lang, "Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture," in *IEEE ICASSP*, vol. 1, 2004, pp. I–577.

[9] J. Nicholson, K. Takahashi, and R. Nakatsu, "Emotion recognition in speech using neural networks," *Neural Computing & Applications*, vol. 9, no. 4, pp. 290–296, 2000.

[10] Y. Zhao, D. Peng, L. Wang, M. Chu, Y. Chen, P. Yu, and J. Guo, "Constructing stylistic synthesis databases from audio books." in *INTERSPEECH*, 2006.

[11] F. Eyben, S. Buchholz, and N. Braunschweiler, "Unsupervised clustering of emotion and voice styles for expressive TTS," in *IEEE ICASSP*, 2012, pp. 4009–4012.

[12] M. Mehrabani and J. H. Hansen, "Singing speaker clustering based on subspace learning in the GMM mean supervector space," *Speech Communications*, vol. 55, no. 5, pp. 653–666, 2013.

[13] S. Ioffe, "Probabilistic linear discriminant analysis," *Computer Vision–ECCV*, pp. 531–542, 2006.

[14] S. Prince and J. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *IEEE International Conference on Computer Vision*, 2007, pp. 1–8.

[15] W. Campbell, D. Sturim, and D. Reynolds, "Support vector machines using gmm supervectors for speaker verification," *IEEE Signal Processing Letters*, vol. 13, no. 5, pp. 308–311, 2006.

[16] F. Castaldo, D. Colibro, E. Dalmasso, P. Laface, and C. Vair, "Acoustic language identification using fast discriminative training," in *Interspeech*, vol. 7, 2007, pp. 346–349.

[17] L. Chen and Y. Yang, "Applying emotional factor analysis and i-vector to emotional speaker recognition," in *Biometric Recognition*. Springer, 2011, pp. 174–179.

[18] H. Hu, M.-X. Xu, and W. Wu, "GMM supervector based SVM with spectral features for speech emotion recognition," in *IEEE ICASSP*, vol. 4, 2007, pp. 413–416.

[19] C. Bishop, *Neural networks for pattern recognition*. Oxford University Press, 1995.

[20] N. Sato and Y. Obuchi, "Emotion recognition using mel-frequency cepstral coefficients," *Information and Media Technologies*, vol. 2, no. 3, pp. 835–848, 2007.

[21] J. Liscombe, J. Venditti, and J. Hirschberg, "Classifying subject ratings of emotional speech using acoustic features," in *Eurospeech*, 2003.

[22] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of german emotional speech," in *Interspeech*, 2005, pp. 1517–1520.

[23] J. Ward Jr, "Hierarchical grouping to optimize an objective function," *Journal of the American statistical association*, pp. 236–244, 1963.