A PERSONALIZED EMOTION RECOGNITION SYSTEM USING AN UNSUPERVISED FEATURE ADAPTATION SCHEME

Tauhidur Rahman and Carlos Busso

Multimodal Signal Processing (MSP)
Department of Electrical Engineering, The University of Texas at Dallas
txr100020@utdallas.edu, busso@utdallas.edu

ABSTRACT

A personalized emotion recognition system aims to tune the model to recognize the expressive behaviors of a targeted person. Such a system can play an important role in various domains including call center and health care applications. Adapting any general emotion recognition system for a particular individual requires speech samples and prior knowledge about their emotional content. These assumptions constrain the use of these techniques in many real scenarios in which no annotated data is available to train or adapt the models. To address this problem, this paper introduces an unsupervised feature adaptation scheme that aims to reduce the mismatch between the acoustic features used to train the system and the acoustic features extracted from the unknown targeted speaker. The adaptation scheme uses our recently proposed *iterative feature normalization* (IFN) framework. An emotion detection system is trained with the IEMOCAP database. For testing, a database was created by downloading videos from a video-sharing website, containing various interviews from a targeted subject (1.5 hours). The detection system is used to identify emotional speech with and without the proposed feature adaptation scheme. The experimental results indicate that the proposed approach improves the unweighted accuracy from 50.8% to 70.0%.

Index Terms— Personalized emotion recognition, front-end feature normalization, feature adaptation.

1. INTRODUCTION

One of the main challenges in emotion recognition is that models and classifiers do not generalize when there is a mismatch between training and testing conditions. This problem is particularly observed when the data from a subject in the testing set is not included during training. In fact, previous studies have shown that speaker dependent classifiers yield higher performance than speaker independent classifiers [1]. This finding suggests that the expression of emotions presents speaker dependencies. Although there are patterns that are observed across speakers (e.g., higher F0 and energy values for angry sentences), these features are not enough to build accurate emotion recognition systems. The problem is that practical applications require robust classifiers that are able to generalize for the expressive speech of unseen speakers. The paper addresses the problem of tailoring an emotion recognition system to a targeted user (i.e., a personalized emotion recognition system). An intriguing approach is the use of feature and/or model adaptation for emotion recognition [2]. A challenge in model adaptation framework is that it requires data with emotional label from the user, which restricts their use in many practical applications.

In this context, this paper introduces an unsupervised feature normalization scheme to adapt an emotion detection system to a targeted user. The proposed adaptation system aims to minimize the mismatch between acoustic features used during training and the acoustic features extracted from the targeted speaker used during testing. The approach relies on our recently proposed *iterative feature normalization(IFN)* scheme [3], which was designed to reduce the speaker variability, while still preserving the signal information critical to discriminate between emotional states. The *Interactive Emotional Dyadic Motion Capture* (IEMOCAP) database is used to build a baseline emotion detection system based on Support Vector Machine (SVM). Videos from a popular video-sharing website, containing various interviews from a targeted subject are downloaded and 1.5 hours of speech from the targeted speaker is extracted. The speech is properly segmented and evaluated in order to test the feature adaptation scheme in real conditions. The experimental results indicate that the proposed unsupervised feature adaptation scheme improves the unweighted accuracy of the emotion detection system from 50.8% to 70.0%.

2. METHODOLOGY

Figure 1 compares a conventional approach (top) with the proposed scheme (bottom). The difference is the unsupervised feature adaptation scheme, which minimizes the mismatch between training and testing conditions in the feature space (Sec. 3). Both cases require an emotion detection system (neutral versus emotional speech), which is described in this section.

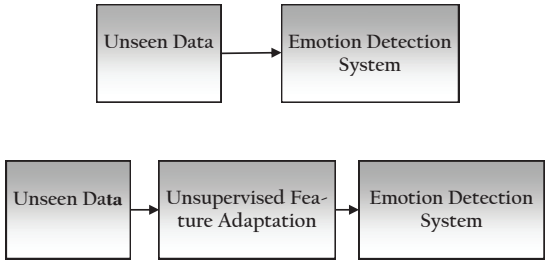

Fig. 1. Baseline approach without feature normalization (top), and proposed unsupervised feature adaptation scheme for emotion detection (bottom).

2.1. IEMOCAP Database

The *Interactive Emotional Dyadic Motion Capture* (IEMOCAP) database is used to train the emotion detection system [4]. Ten trained actors were recorded in five dyadic interactions using scripts and improvisations. The plays and scenarios were selected to evoke sadness, happiness, anger and frustration. As a result of the dialog between the actors, other emotions such as surprise and excitement were also elicited. The corpus contains approximately twelve hours of data, which was manually segmented and transcribed at the turn level. The emotional content of each turn was annotated by three different evaluators with categorical labels such as neutral, sadness, happiness, anger and frustration. The turns were also annotated by two evaluators with attribute-based labels corresponding to valence [1-negative, 5-positive], activation [1-calm, 5-excited] and dominance [1-weak, 5-strong]). For this study, the emotional primitive values were mapped into the range [-1, +1].

In this study, we are interested in a binary emotion detection problem (neutral versus emotional speech). To build an emotion classifier, the corpus is split into neutral and emotional classes. We study 1109 sentences labeled as neutral and 3777 sentences labeled with other discrete emotional categories. To increase the reliability of the neutral labels, we established a circle centered at the origin of the coordinate system defined by the activation-valence space. Only neutral samples lying in this inner circle were selected. The radius of this circle was set to 0.5, which includes approximately 75% of the neutral sentences. Similarly, among the sentences labeled with emotional labels, we only selected those in which their activation-valence values lie outside of the circle centered at the origin with radius 0.6. These values were empirically chosen to capture clearer emotional content.

2.2. Feature Set

The openSMILE toolkit with the INTERSPEECH 2009 Emotion Challenge feature set is used to extract a set of common acoustic and prosodic features. The set contains 384 sentence level features [5].

Table 1. Emotion detection results on the IEMOCAP database. The table compare the accuracy of the systems implemented with different feature normalization schemes.

Normalization	Accuracy(%)
Without normalization	69.8%
Iterative feature normalization	71.8%
Perfect normalization	72.8%

2.3. Emotion Detection System

A linear kernel Support Vector Machine (SVM) with sequential Minimal Optimization (SMO) is used as our emotion detector. This classifier has been successfully used in other paralinguistic recognition problems [6]. The SVM is trained and tested with the WEKA data mining toolkit using all 384 features. The complexity parameter of the classifier is empirically set to 0.1. To estimate the performace of the system, this study uses leave-one-speaker-out 10-fold cross validation. In each fold, samples from the emotional class are randomly selected to match the number of samples in the neutral class. This process is performed in the training and testing partitions to compensate unbalanced classes (chance is 50%). This random selection is repeated 100 times for each fold. Table 1 shows the average performance of this emotion detection system. Without any normalization, the classifier reaches the accuracy of 69.8%. Notice that IEMOCAP is a very challenging spontaneous database, with samples conveying ambiguous emotional content [7].

3. UNSUPERVISED FEATURE ADAPTATION

This paper proposes the use of feature normalization to adapt an emotion detection system to a particular individual (Fig. 1). The purpose of normalizing acoustic features is to reduce speaker variability, while preserving the discrimination between emotional classes. This goal can be achieved by reducing the differences observed between the acoustic features used to train and test the emotion detection system. In this context, we proposed the use of the *iterative feature normalization* (IFN) approach [3] as a front-end.

We have shown that global normalization (i.e., estimated normalization parameters across the entire corpus including neutral and emotional samples) affects the emotion discrimination of the features [3]. A better approach consists in reducing the differences observed across the neutral subset of the speakers (i.e., forcing that the properties in acoustic features derived from neutral speech are similar across speakers) [8]. For each speaker, the normalization parameters are estimated from his/her neutral samples, and then applied to his/her entire data (both neutral and emotional partitions). Notice that this speaker dependent normalization scheme preserves the emotional variability in the features.

We have proposed the IFN framework, which implements

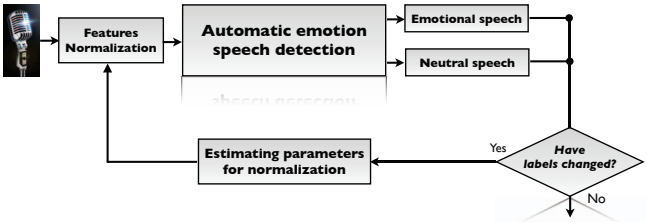

Fig. 2. Overview of the iterative feature normalization (IFN) approach proposed by Busso et al. [3].

the aforementioned ideas (Fig. 2) [3]. This unsupervised front-end scheme iteratively estimates the normalizing parameters from an unseen speaker. Therefore, it is a suitable framework for our personalized emotion detection system. In the IFN scheme, an emotion detection system is used to iteratively identify neutral speech of the unseen speaker. Then, it estimates the normalization parameters using only this subset (relying on the detected labels). These normalization parameters are then applied to the entire data, including the emotional samples. This detection-normalization approach is iteratively repeated. The performance of the emotion detection algorithm is expected to improve, leading to better normalization parameters. The process is repeated until the percentage of the samples that change their emotional labels is below a given threshold (e.g., 5%). We have shown that this normalization approach increases the performance of an emotion detection system [3]. Unlike our previous work, the IFN approach is implemented using Z-normalization for each feature (i.e., the normalization parameters are the mean and standard deviations of the sentence level features).

Table 1 shows the performance of the emotion detection system presented in Sec. 2.3 with the IFN approach as front-end. The accuracy increases compared to the case without feature normalization (71.8%). If the neutral samples were perfectly detected during the iterative process (i.e., using the emotional labels to normalize the features), the performance of the emotion detection system reaches 72.8% (third row of Table 2.3), which is the best performance for this data reported in the paper. Notice that the IFN approach almost reaches the accuracy achieved when the features are normalized with the optimal coefficients.

4. PERSONALIZED EMOTION DETECTION

The unsupervised feature adaptation scheme is tested with a realistic database containing speech from a single speaker during various uncontrolled recordings. This section describes this data and the experimental results.

4.1. Personalized Data Collection

Ninety minutes of audiovisual data from one speaker was downloaded from a video-sharing website. The videos correspond to talks and interviews from a recognized actress in Hollywood. This work only considers the speech channel. To simplify the task, noisy and overlapped segments were carefully manually removed. The speech of the targeted speaker is automatically segmented into 5 second long speech files with no overlap, resulting in 837 segments. All the files are converted to have the same sampling frequency (16Khz, single channel with 16 bit of sample size).

The motivation behind collecting data from a particular individual is to investigate whether the proposed feature adaptation scheme can compensate the speaker variability from the unseen speaker. Notice that the data incorporates many real challenging conditions including different environmental conditions, recordings, different ages from the actress (from 15 to 30 year old) and channel effect. Notice that an emotion detection system should cope with these conditions, which are not observed in controlled databases.

The speech files were perceptually evaluated. For this purpose, three Graduate students without any hearing problem evaluated the emotional content of each signal. The subjects were asked to evaluate the files using a *graphical user interface* (GUI) with a slider bar designed for this evaluation (0-neutral, 1-emotional). The GUI gives the subjects the option to return to any previous file to adjust the rating. The average value across evaluators is used as ground truth. If the rating is lower than 0.5, the file is considered as neutral. Otherwise, the file is considered as emotional. As expected, the database is emotionally unbalanced with more neutral (738) than emotional (99) samples.

Table 2. Performance of the emotion detection system with and without the proposed unsupervised feature normalization scheme. Results are reported in terms of Weighted Accuracy (WA) and Unweighted Accuracy (UA).

Emotion detection system	WA(%)	UA(%)
Without Normalization	36.3%	50.8%
Unsupervised Feature Adaptation	80.3%	70.0%

4.2. Experimental Results

Table 2 shows the performance of the emotion detection system without normalization and with the proposed unsupervised feature normalization scheme. The performance is presented in terms of unweighted (UA) and weighted (WA) accuracies. UA is particularly challenging given the unbalanced corpus. The UA of the feature adaptation scheme with the proposed emotion detection system is 70.0% , which is significantly higher than the performance without any normalization (50.8%). The improvement in WA is even higher, increasing from 36.3% to 80.3%.

Figure 3 shows the histograms of the predicted classes by the detection system without (Fig. 3(a)) and with (Fig. 3(b)) the proposed feature adaptation scheme. The x-axis corresponds to the averaged values assigned to speech samples. In these figures, light gray is used for the histograms of sam-

Table 3. Confusion matrices of the emotion detection systems without (a) and with (b) the proposed unsupervised feature normalization scheme (Emo=emotional, Neu= neutral).

(a) Baseline

	Neu	Emo
Neu	235	503
Emo	30	69

(b) Proposed approach

	Neu	Emo
Neu	616	122
Emo	43	56

ples detected as neutral, and dark gray is used for samples detected as emotional. Table 3 provides the confusion matrices for both systems. Figure 3 and Table 3 shows that the precision of the emotion detection system without normalization is very low (36.3%). Although the recall for emotional class is higher (70%), the system is achieving this performance at the expense of misclassifying 70% of neutral samples. Even though the recall of the emotional samples is lower than the baseline system (56.6%), the proposed adaptation scheme significantly increases the precision of the system (31.5%). It mislabels only 16% of the neutral samples, while correctly predicting 56% of the emotional samples. The results illustrate the negative affect that channel and speaker mismatches have on the performance of an emotion detection system. The proposed unsupervised feature adaptation scheme is able to compensate these differences, up to some extent, achieving higher unweighted and weighted accuracies than the baseline system.

5. CONCLUSIONS

This paper describes our effort to implement a personalized emotion recognition system using an unsupervised feature adaptation scheme. The proposed front-end framework is able to reduce the mismatches in the training and testing conditions by iteratively estimating and applying the normalization parameters to acoustic features extracted from an unseen speaker.

The benefits of using this feature adaptation scheme is demonstrated in controlled and uncontrolled recording conditions. The results on the IEMOCAP database indicated that the accuracy of the proposed system is 2% (absolute) higher than the one achieved by the baseline without the feature adaptation scheme. The results on the uncontrolled recordings (i.e., speech downloaded from a video-sharing website) revealed that the feature adaptation scheme significantly improved the unweighted and weighted accuracies of the emotion detection system.

Figure 3(b) and Table 3 show that there are samples that are misclassified even after feature adaptation. Toward reducing the detection errors, we are currently exploring model adaptation techniques that can be coupled with the proposed front-end unsupervised feature normalization scheme. We expect that a combination of different adaptation techniques will improve the performance and robustness of the proposed personalized emotion recognition framework.

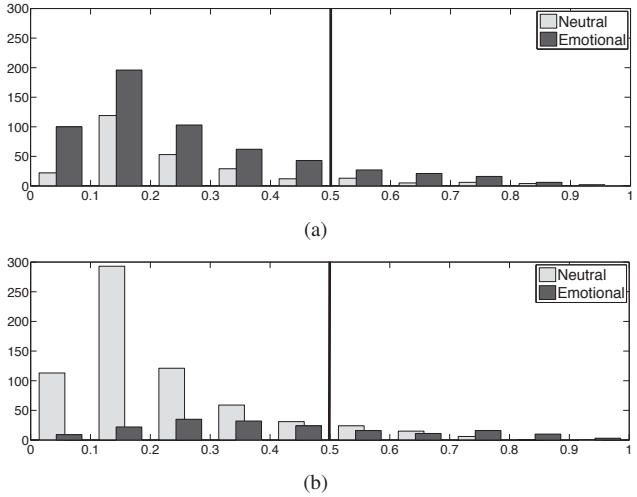

Fig. 3. Histograms with the emotional classes predicted by (a) baseline emotion detection system and by (b) emotion detection system with unsupervised feature adaptation.

6. REFERENCES

[1] A. Austermann, N. Esau, L. Kleinjohann, and B. Kleinjohann, "Fuzzy emotion recognition in natural speech dialogue," in *Robot and Human Interactive Communication, 2005. ROMAN 2005. IEEE International Workshop on*, aug. 2005, pp. 317 – 322.

[2] J.-B. Kim, J.-S. Park, and Y.-H. Oh, "On-line speaker adaptation based emotion recognition using incremental emotional information," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP-2011)*, may 2011, pp. 4948–4951.

[3] C. Busso, A. Metallinou, and S. Narayanan, "Iterative feature normalization for emotional speech detection," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2011)*, Prague, Czech Republic, May 2011, pp. 5692–5695.

[4] C. Busso, M. Bulut, C.C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J.N. Chang, S. Lee, and S.S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Journal of Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, December 2008.

[5] B. Schuller, S. Steidl, and A. Batliner, "The INTERSPEECH 2009 emotion challenge," in *Interspeech 2009 - Eurospeech*, Brighton, UK, September 2009.

[6] T. Rahman, S. Mariooryad, S. Keshavamurthy, G. Liu, J.H.L. Hansen, and C. Busso, "Detecting sleepiness by fusing classifiers trained with novel acoustic features," in *12th Annual Conference of the International Speech Communication Association (Interspeech'2011)*, Florence, Italy, August 2011.

[7] E. Mower, A. Metallinou, C.-C. Lee, A. Kazemzadeh, C. Busso, S. Lee, and S.S. Narayanan, "Interpreting ambiguous emotional expressions," in *International Conference on Affective Computing and Intelligent Interaction (ACII 2009)*, Amsterdam, The Netherlands, September 2009.

[8] C. Busso, S. Lee, and S.S. Narayanan, "Analysis of emotionally salient aspects of fundamental frequency for emotion detection," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 17, no. 4, pp. 582–596, May 2009.