SPEAKER VARIABILITY IN EMOTION RECOGNITION – AN ADAPTATION BASED APPROACH

Ni Ding, Vidhyasaharan Sethu, Julien Epps, Eliathamby Ambikairajah

The School of Electrical Engineering and Telecommunications,
The University of New South Wales, Sydney NSW 2052, Australia

ABSTRACT

None of the features commonly utilised in automatic emotion classification systems completely disassociate emotion-specific information from speaker-specific information. Consequently, this speaker-specific variability adversely affects the performance of the emotion classification system and in existing systems is frequently mitigated by some form of speaker normalisation. Speaker adaptation offers an alternative to normalisation and this paper proposes a novel bootstrapping technique which involves selecting appropriate initial models from a large training pool, prior to speaker adaptation of emotion models in the context of GMM based emotion classification as an alternative to speaker normalisation. Evaluations on the LDC Emotional Prosody and the FAU Aibo corpora reveal that an emotion classification system based on the proposed bootstrapping method outperforms systems based on speaker normalisation as long as a small amount of labelled adaptation data is available. It also outperforms speaker adaption from common initial models estimated from all training speakers.

Index Terms— Speaker adaptation, emotion classification, speaker normalisation, bootstrapping

1. INTRODUCTION

Human speech is a rich source of information. Apart from the actual linguistic component comprising a sequence of phonetic units conveying a message, speech also contains paralinguistic cues such as those specific to the speakers and those that express emotions. Systems that recognize these cues, such as emotion classification systems, are therefore generally designed to function in two broad stages: a front-end that extracts features intended to be characteristic of these paralinguistic cues and a back-end that recognises the emotion based on these features. Ideally, the only source of variability in the extracted features would be due to differences in the emotions being expressed. However, variability in features extracted from speech arises due to numerous other reasons as well, including linguistic content (due to differences between what is being said) and speaker identity (due to differences between who is saying it). These additional sources of variability in turn degrade classification performance [1].

Back-ends based on Gaussian mixture models (GMMs), while conceptually straightforward, have been shown to be extremely versatile and powerful in various speech based classification systems including emotion classification [2]. Since GMMs model the probability distributions of each emotion independent of other sources of variability, these sources affect all the models indiscriminately. While linguistic and speaker variability would most probably be the two most significant influences on an emotion classification system, they may not affect performance in the same way. It has been suggested that speaker variability is a more significant issue [3].

Typically some sort of speaker normalisation is utilised to reduce speaker specific variability in the features prior to modelling and testing, and this has been shown to result in significant gains in terms of classification accuracy [4]. A range of techniques include mean normalisation, cumulative distribution mapping [4] and joint factor analysis [5] have been employed for this purpose. All of these techniques attempt to remove the effect of speaker variability in the feature (or model) space. However, none of these methods for separating the effects of speaker variability from that of emotion variability are completely accurate and hence in practical scenarios result in some unintentional loss of emotion-specific information or in some residual speaker-specific information remaining. This observation is supported by the observation that speaker independent emotion classification systems do not perform as well as speaker dependent ones even if they incorporate speaker normalisation. Rather than normalise the features in order to minimise the mismatch between trained models and the test speaker, an alternative approach is to adapt the emotion-specific models of the back-end towards a target speaker. This is potentially superior to normalisation since it does not remove any information from the feature space. An adaptation approach can adapt initial emotion models estimated from training speakers' data to match the target speaker. Typically, adaptation such is performed using an initial model trained on multiple speakers [6], however such a model would already be affected by speaker variability.

This paper investigates a speaker adaptation approach where the initial emotion models are chosen from a large set of speaker specific emotion models (from a single speaker so that the models are not affected by speaker variability) and compares it with a speaker normalisation one for a GMM based emotion classification system.

2. EMOTION RECOGNITION SYSTEM

2.1. Front-End

Since the aim of the study is to compare speaker adaptation with speaker normalisation, the choice of features is not as critical as the fact that the same features are used in all experiments. The conventionally employed MFCCs were chosen as features in all experiments. The front-end used 20ms frames with Hamming window and 50% overlap to extract 12 MFCCs per frame. This was followed by a VAD to discard unvoiced frames and only features from voiced frames in each utterance were used.

2.2. Speaker Normalisation

Feature warping, or cumulative distribution mapping, is a technique that maps each feature to a predetermined distribution, originally suggested as a method to provide robustness against channel mismatch and nonlinear noise effect. Previously, we have used a modified feature warping technique as a means of speaker normalisation [4] and have utilised it for the same purpose in some of the experiments reported in this paper.

2.3. Back-End

Almost all current automatic emotion classification systems utilise statistical machine learning back-ends such as Gaussian mixture

models [2], hidden Markov models [7], support vector machines [8], neural networks [9]. While some studies that suggest that a multi-stage approach, akin to those used in speaker verification systems, may give good results [2, 5], there are still no studies that strongly indicate any one approach is superior to others.

Given that the aim of the experiments reported in this paper is to compare a speaker adaptation approach to a speaker normalisation one, a Gaussian mixture model (GMM) based back-end was chosen. The main reasons for this choice were that MAP adaptation of GMMs is well established [10] and GMM-UBM approaches have been used in many speech based classification systems.

2.4. Databases

The English LDC Emotional Prosody speech corpus [11] and the German FAU Aibo Emotion Corpus [12] were used in the experiments reported in this paper. The LDC corpus consists of speech from professional actors trying to express emotions while reading short phrases consisting of dates and numbers. The entire database consists of 7 actors expressing 15 emotions for around 10 utterances each. Data from five emotions, namely anger, sadness, happiness, boredom and neutral (no emotion) were used in all experiments, set up as 5-class classification problems. For speaker-dependent systems, 70% of all utterances from each speaker for each of the 5 emotions were used as training data and the remaining 30% as test data. The systems were trained and tested 7 times (once per speaker) and the seven results averaged to obtain the final accuracies. In all other systems (speaker-independent), a 7-fold leave-one-out type cross validation was carried out with data from 6 speakers used as training and the data from the 7^{th} split in a 2:8 ratio for use as adaptation and evaluation data in each fold. The results of the seven folds were averaged to obtain final accuracies.

The German Aibo corpus consists of spontaneous emotionally coloured children's speech with recordings of 51 German children aged between 10 and 13 from two different schools. This database was used in the INTERSPEECH-09 Emotion Challenge [13] for a 5-class classification task, and data from children of one school were used as the training set and data from children of the other one as the test set. The unweighted average recall (UAR) was used as the evaluation metric in the challenge. In the experiments reported in this paper, the same 5-class task was retained along with UAR as the evaluation metric. In addition to the UAR, the paper also reports the weighted average recall (WAR). The training set was also identical to the one defined for the emotion challenge. The test set was however, split further into adaptation and evaluation sets. 10% of all speech chunks from each speaker in the test set were allocated to the adaptation set and the remaining 90% to the evaluation set. Further, unlike the LDC corpus, the data from the Aibo corpus were not used in speaker-dependent systems since the initial training and test set division did not allow for common speakers. All speaker-independent systems reported in this paper that were tested on the Aibo corpus were evaluated only on the evaluation set (90% of the original test set), regardless of whether the system used the adaptation data or not. This was done to ensure the results could be compared to each other.

2.5. Baseline Speaker-Independent System

The speaker-independent (S-IND) systems used in the experiments reported in this paper used a GMM to model the feature distributions of each emotion. These GMMs were trained on labelled data from multiple speakers from the training set and tested on data from speakers not included in the training set. When carried out, speaker normalisation was applied to the features prior to training and testing, and no adaptation was performed. Table 1 reports the performance of the speaker-independent system with and without speaker normalisation when evaluated on the LDC database while Table 2 reports the performance obtained when evaluated on the Aibo corpus. The small drop in UAR when normalisation is carried out is most likely due to the unbalanced nature of the database.

2.6. Baseline Speaker-Dependent System

The configuration of the speaker-dependent (S-DEP) system used is almost identical to the speaker-independent system, with one difference. Namely, the training and test data came from the same speaker. The two sets (train and test) were still distinct in that no utterances occurred in both. Table 1 reports the classification accuracies obtained when the speaker-dependent system was evaluated on the LDC corpus. Given that speaker normalisation is redundant in a speaker-dependent system, it is not surprising to see that its use only results in a small drop in performance.

Table 1: Classification accuracies for baseline systems (LDC corpus).

System	Overall Classification Accuracy (%)	
	with normalisation	without normalisation
Speaker-Dependent	81.9 %	83.8 %
Speaker-Independent	55.1 %	51.6 %

Table 2: Classification accuracies for baseline systems (Aibo corpus).

System	Recall (%)			
	with normalisation		without normalisation	
Speaker-Independent	UA	WA	UA	WA
	35.7 %	27.7 %	37.7%	35.5 %

3. SPEAKER ADAPTATION

3.1. Motivation

Comparing the classification accuracies of the speaker-dependent system to those of the speaker-independent system (without normalisation) in Table 1 makes it clear that speaker variability plays a significant role in emotion classification systems and needs to be addressed. If the feature vectors corresponding to different emotions can be thought of as occupying different regions of the feature space (with the amount of overlap being proportional to the confusability between the overlapping emotions), the distribution of these regions is speaker-specific to some degree. Therefore models trained on data from one (or more) speaker(s) may not coincide with the regions corresponding to another and hence result in lower classification accuracy.

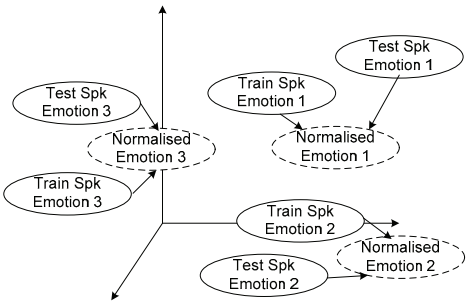

Figure 1: Conceptual illustration of speaker normalisation approach

Speaker normalisation attempts to address this issue by modifying the feature vectors for each speaker in a manner such

that the emotion regions for different speakers align in the modified feature space (dashed ellipses in Figure 1). Speaker adaptation on the other hand uses models trained on one (or more) speaker(s) and attempts to modify the model to match the regions of the target speaker (Figure 2). Typically, these initial models are trained on speech from multiple speakers to create a common set of models that are then adapted to match each of the target speakers. In this scenario however, the initial models are affected by the speaker variability present in the training data and this may affect the accuracy of the adapted models. The proposed technique aims to overcome this issue.

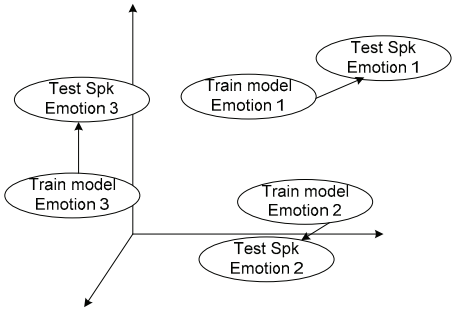

Figure 2: Conceptual illustration of speaker adaptation approach

3.2. Proposed Speaker Bootstrapping

Given sufficient training data, it is possible to train speaker-dependent models (without speaker normalisation) for each emotion, giving rise to one model per speaker per emotion. It is reasonable to suppose that for a target test speaker, some of these models will be a better match than the others. The proposed bootstrapping technique involves using a small amount of adaptation data (labelled) from a test speaker, selecting the closest emotion model set from the trained speaker-dependent model sets and using MAP adaptation [10] to further improve them (with respect to the target speaker) as shown in Figure 3. These adapted models can then be used for emotion recognition.

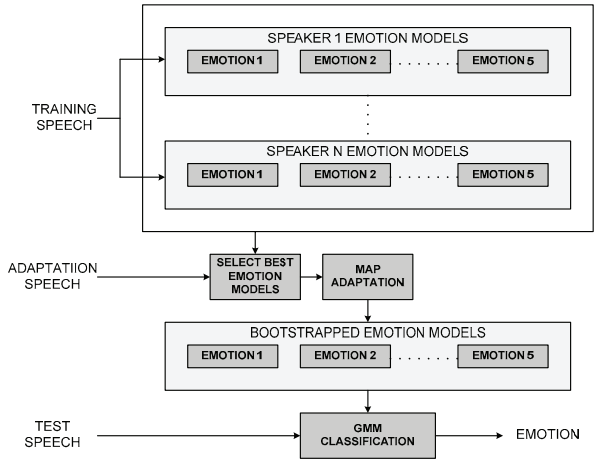

Figure 3: Block diagram outlining the proposed bootstrapping technique

In order to determine the best set of emotion models for adaptation, for each emotion the likelihoods of all speaker-specific models (estimated from training data) given the speech corresponding to that emotion from the adaptation data (for the target speaker) were computed. For emotion then, the trained speaker-dependent model(s) with the highest likelihood were then selected for MAP adaptation. It should be noted that the models selected for adaptation for the each of the different emotions may correspond to a different training speaker. i.e.,

$$\hat{\boldsymbol{\lambda}}_i = \arg\max_{\boldsymbol{\lambda}_i^{(j)}} P\left(\mathbf{X}_i \middle| \boldsymbol{\lambda}_i^{(j)}\right) \quad (1)$$

where, $\hat{\boldsymbol{\lambda}}_i$ is the speaker dependent model corresponding to emotion i chosen as the initial model for adaptation, $\boldsymbol{X}_i$ denotes the adaptation data corresponding to emotion i, and $\boldsymbol{\lambda}_i^{(j)}$ denotes the model of emotion i estimated from data from speaker j.

In some cases, it is possible that none of the training speakers' models match the target speaker particularly well, or more than one training speaker's models provide a good match. In both cases, it may be more suitable to select the n-best matches for each emotion and adapt the hybrid model. Such a hybrid model can be created by training a new GMM using all the data that went into training the n-best selected GMMs for that emotion.

4. EXPERIMENTAL RESULTS

The proposed speaker bootstrapping technique from the single best initial model was first tested on the LDC corpus. Training, adaptation and evaluation datasets were set up as outlined in section 2.4. The classification accuracies obtained are given in Table 3. Bootstrapping from n-best models was not attempted on the LDC corpus due to the small number of speakers available. Comparing the accuracy of the system employing speaker normalisation (55.1 %) to that employing the proposed bootstrapping technique (70.4 %), it is clear that an adaptation based approach is potentially superior to a normalisation based on, even though it still is not as good as a speaker dependent system (83.8 %).

Table 3: Confusion Matrix corresponding to system employing bootstrapping from 1-best model (LDC corpus)

	Neutral	Anger	Sad	Happy	Bored
Neutral	**64.1 %**	0.0 %	10.3 %	5.1 %	20.5 %
Anger	0.0 %	**91.2 %**	1.8 %	5.3 %	1.8 %
Sad	1.8 %	3.5 %	**59.7 %**	12.3 %	22.8 %
Happy	1.6 %	21.9 %	9.4 %	**54. 7 %**	12.5 %
Bored	2.9 %	1.4 %	8.6 %	7.1 %	**80.0 %**
Overall Classification Accuracy = 70.4%					

As an alternative to the proposed bootstrapping technique, a system that adapted speaker independent emotion models towards the target speaker was also implemented and found to have an overall accuracy of 66.2 %. The lower performance relative to the proposed system is most likely because in this case the initial speaker independent models are affected by speaker variability.

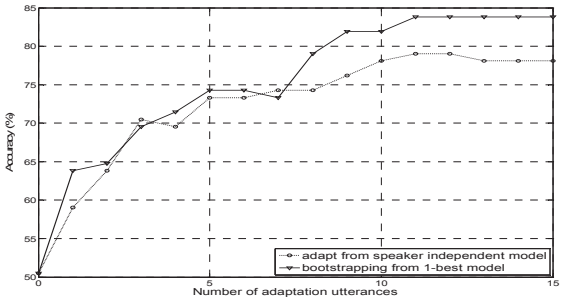

Figure 4: Five-class emotion classification accuracy as a function of the amount of adaptation data (LDC corpus)

Finally, it is reasonable to expect that the performance of the proposed technique will improve with increasing amounts of adaptation data. In order to do this, a different adaptation-evaluation split was utilised, namely 70% of the test speakers data was used for adaptation and the remaining 30% for evaluation. The classification accuracies obtained are graphed in Figure 4.

As expected, the system performance is observed to increase with the size of the adaptation dataset. However, this graph should only be considered indicative of the expected trend and the accuracies not compared directly to the other results since the evaluation dataset is different.

Following the tests on the LDC corpus, the proposed system was validated on the FAU Aibo corpus. The training and test sets were separated as suggested for the INTERSPEECH-09 Emotion Challenge and the evaluation and adaptation sets split as outlined in section 2.4. Bootstrapping from several *n*-best initial models was evaluated, as listed in Table 4. It should be noted that since the INTERSPEECH-09 Emotion Challenge did not define an adaptation dataset, these results cannot be directly compared with the results based on the challenge guidelines. This was unavoidable since the proposed technique requires adaptation data from the target speaker.

Table 4: UARs for systems employing bootstrapping from *n*-best models (Aibo corpus)

	Recall (%)	
n	UA	WA
1	36.6 %	47. 5 %
2	37.7 %	45.1 %
3	**38.5 %**	**44.5 %**
4	37.7 %	44.1 %
5	36.7 %	42.9 %

From the accuracies reported in Table 4, it can be seen that bootstrapping from more than one model can in fact provide an improvement in performance when compared with bootstrapping from just one model. The best performance (in terms of UAR) in this case turns out to be from the system that bootstraps off 3 speakers per emotion. The confusion matrix corresponding to this system is reported in Table 5. Also, as was done for the LDC corpus, a system that adapted speaker-independent emotion models (instead of *n*-best speaker specific models) was evaluated on the Aibo corpus and was found to exhibit an UAR of 37.4% (WAR of 42.1%). While the system based on the proposed technique exhibits a much smaller improvement over the one based on speaker normalisation when testes on the Aibo corpus as opposed to the LDC corpus, this is probably due to the imbalance in the adaptation data, which contains more neutral speech than speech representing other emotion classes. Further, this imbalance affects the speaker normalisation process even more severely than it does the proposed bootstrapping technique. It should finally be noted that the proposed technique is limited by the fact that a small amount of emotionally labelled adaptation data from the target speaker is required.

Table 5: Confusion matrix corresponding to system employing bootstrapping from 3-best models (Aibo Corpus)

	Angry	Emphatic	Neutral	Positive	Rest
Angry	**47.5 %**	31.5 %	10.8 %	1.3 %	8.9 %
Emphatic	21.7 %	**59.3 %**	12.1 %	1.4 %	5.5 %
Neutral	18.1 %	26.2 %	**43.4 %**	2.6 %	9.8 %
Positive	9.9 %	6.6 %	44.0 %	**24.7 %**	14.8 %
Rest	17.5 %	24.0 %	36.9 %	4.0 %	**17.7 %**

6. CONCLUSION

This paper has shown that speaker adaptation is a viable and potentially superior alternative to speaker normalisation in the context of emotion classification. While a normalisation based approach, by its nature, always results in some information being lost, an adaptation based approach faces no such constraint and its performance is bounded only by that of a speaker-dependent system. In fact, given sufficient adaptation data, the system employing the proposed bootstrapping technique seems to converge to a speaker dependent system, even though this situation would rarely ever occur in practice. The experimental results included in this paper show that systems based on the proposed technique consistently outperform those based on normalisation.

7. ACKNOWLEDGEMENT

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

8. REFERENCES

[1] A. Batliner and R. Huber, "Speaker Characteristics and Emotion Classification
Speaker Classification I." vol. 4343, C. Müller, Ed., ed: Springer Berlin / Heidelberg, 2007, pp. 138-151.

[2] I. Luengo, E. Navas, and I. Hernaez, "Combining spectral and prosodic information for emotion recognition in the Interspeech 2009 Emotion Challenge," in *INTERSPEECH-2009*, 2009, pp. 332-335.

[3] V. Sethu, E. Ambikairajah, and J. Epps, "Phonetic and Speaker Variations in Automatic Emotion Classification," in *INTERSPEECH-2008*, 2008, pp. 617-620.

[4] V. Sethu, E. Ambikairajah, and J. Epps, "Speaker Normalisation for Speech-Based Emotion Detection," in *Digital Signal Processing, 2007 15th International Conference on*, 2007, pp. 611-614.

[5] M. Kockmann, L. Burget, and J. Cernocky, "Brno University of Technology System for Interspeech 2009 Emotion Challenge," in *INTERSPEECH-2009*, 2009, pp. 348-351.

[6] K. Jae-Bok, P. Jeong-Sik, and O. Yung-Hwan, "On-line speaker adaptation based emotion recognition using incremental emotional information," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, 2011, pp. 4948-4951.

[7] R. Huang and C. Ma, "Toward A Speaker-Independent Real-Time Affect Detection System," in *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, 2006, pp. 1204-1207.

[8] D. Bitouk, R. Verma, and A. Nenkova, "Class-level spectral features for emotion recognition," *Speech Communication,* vol. 52, pp. 613-625, 2010.

[9] M. W. Bhatti, W. Yongjin, and G. Ling, "A neural network approach for human emotion recognition in speech," in *Circuits and Systems, 2004. ISCAS '04. Proceedings of the 2004 International Symposium on*, 2004, pp. II-181-4 Vol.2.

[10] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models," *Digital Signal Processing,* vol. 10, pp. 19-41, 2000.

[11] M. Liberman, K. Davis, M. Grossman, N. Martey, and J. Bell. [Online]. Available: http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2002S28

[12] S. Steidl, *Automatic Classification of Emotion-Related User States in Spontaneous Children's Speech*. Berlin: Logos Verlag, 2009.

[13] B. Schuller, S. Steidl, and A. Batliner, "The INTERSPEECH 2009 Emotion Challenge," presented at the INTERSPEECH-2009, Brighton, UK, 2009.