SPEAKER VARIABILITY IN SPEECH BASED EMOTION MODELS – ANALYSIS AND NORMALISATION

Vidhyasaharan Sethu, Julien Epps, Eliathamby Ambikairajah

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

ABSTRACT

All features commonly utilised in speech based emotion classification systems capture both emotion-specific information and speaker-specific information. This paper proposes a novel method to gauge the effect of speaker-specific information on emotion modelling based on two measures: a Monte Carlo approximation to KL divergence and an estimate of feature variability based on diagonal covariance matrices. In addition, a novel speaker normalisation technique based on joint factor analysis is also proposed. This method is analogous to channel compensation in speaker verification systems, with one significant extension. The model domain compensation is mapped back to frame-level features, allowing for use in a wider range of emotion classification frameworks and in conjuncture with other normalisation techniques. Preliminary evaluations on the IEMOCAP database suggests that the proposed technique improves the performance of GMM based classification systems based on widely employed features such as pitch, MFCCs and deltas.

Index Terms— KL divergence, joint factor analysis, speaker normalisation, emotion classification

1. INTRODUCTION

Systems that recognise paralinguistic cues based on speech, such as emotion classification systems, generally operate in two broad stages. A front-end that extracts features characteristic of the paralinguistic information of interest and a back-end that makes classification decisions based on these features. Almost universally features tend to be vector representations of speech signals and back-end classification decisions are based on differences in the statistical properties of the distributions of the feature vectors. Consequently the performance of speech based emotion classification systems depends on two factors, namely, the degree to which the underlying statistical properties of the feature vector distributions estimated from speech corresponding to different emotions differ and the accuracy with which these differences can be modelled by the back-end. The first factor determines an upper bound on the classification accuracy of any emotion classification system given a set of features, while the second factor leads to differences in the classification accuracies of the different systems.

Ideally, the statistical properties of feature vector distributions would vary significantly between different emotions (herein referred to as emotional variability) and not vary due to any other reason. However, in reality, they also vary significantly due to differences between different speakers (speaker variability), due to differences in linguistic content (phonetic variability) and also differences in other paralinguistic cues. These additional sources of variability in turn affect the 'classification rules' inferred by the back-end and degrade classification performance [1-3]. While phonetic and speaker variability would most probably be the two most significant influences on an emotion classification system, it has been suggested that speaker variability is a more significant issue in many commonly utilised features [4].

Approaches to compensate for speaker variability in emotion classification systems can be broadly categorised into those that explicitly personalise the systems towards a target speaker or those that alter the feature vectors or models of their distributions to minimise the effect of speaker variability on them. The former category includes systems with back-ends trained exclusively on data from the target speaker [5] and those with a generic back-end that is then suitably adapted towards target speakers [6, 7]. The latter category consists of techniques, referred to herein as speaker normalisation techniques, which aim to reduce speaker variability either in the feature domain or in the domain of models of feature distributions. Feature domain and model domain techniques are both designed to minimise the effect of speaker variability on the statistical properties of the feature vector distributions. Specifically, feature domain techniques modify feature vectors directly [8-10] while model domain techniques modify representation of models (such as supervectors) [11, 12]. In almost all cases, the speaker normalisation techniques have shown improved performance (to varying degrees) but there has been little work analysing in detail how speaker variability affects the feature distributions in the first place. Such analyses may help motivate speaker normalisation techniques designed to improve them. This paper attempts such an analysis and presents a normalisation technique leading from the analysis.

2. RELATION TO PRIOR WORK

While studies have shown that speaker variability has a negative impact on the performance of emotion classification systems [2] and have proposed speaker normalisation techniques that improve performance [8-12], there is a dearth of analyses on how this speaker variability manifests itself. This paper reports a novel investigation of both the nature and the extent of the effect of speaker variability on feature vector distributions. Further, based on this analysis, it proposes a novel speaker normalisation approach based on joint factor analysis (JFA) to compensate for some of the effects identified (and roughly quantified). A Speaker ID system adapted for emotion classification [13] had included JFA (applied in the model domain on supervectors) as part of the framework but was used with restricted modelling ability (small number of parameters) and concluded the improvements were negligible. The technique proposed in this paper differs from speaker ID type approaches [14-16] by applying a model domain JFA based normalisation and extending it by mapping the compensation back to the feature domain. Such an approach also allows it to be used in a wider range of systems.

3. DATABASE

The IEMOCAP (Interactive emotional dyadic motion capture) database [17] was used in all the work reported in this paper. The database consists of audio-visual recordings of five sessions of

dyadic mixed-gender pairs of actors in either improvised affective scenarios or scripted scenarios. The recorded dialogues have been manually segmented into utterances, each of which have been categorically annotated with emotion. In the work reported in this paper, the manually segmented audio recordings from all 10 speakers associated with the emotional categorical labels anger, happiness, excitation, neutrality and sadness were used. Further, the classes of happiness and excitation were merged into a single class (happiness) to create a 4 emotional class scenario as in [11].

Half the utterances from each speaker, corresponding to each of the 4 emotional classes, were used as a training set and the other half as a test set in all experiments. This approach was taken instead of the somewhat more common leave-one-out cross-fold validation since the focus of the paper is on speaker variability and training data from all 10 speakers was used to learn the normalisation parameters. It should be noted that the all classification experiments reported were still carried out in a speaker-independent manner using data from all the speakers together without identifying individual speakers in both training and testing phases. It has been suggested that JFA parameters, such as those in the proposed technique, are not estimated accurately with small amounts of data [13] and IEMOCAP is one of the few publically available databases that contains a reasonable large amount of speech data from each speaker for each emotion.

4. SPEAKER AND EMOTIONAL VARIABILITY

This section presents a novel analysis of the effect of emotion variability and speaker variability on a feature space. Specifically, it compares models of probability distributions of features estimated from speech corresponding to different speakers and different emotions. Gaussian mixture models (GMMs) are used to model probability distributions on the feature space and symmetric KL divergence is used as an estimate of dissimilarity between models. All features were extracted from 20ms frames (except pitch) with shifts between consecutive frames. Only voiced frames (voicing determined by the pitch extraction algorithm [18]) were used in all analyses and by all classification systems.

4.1 Symmetric KL Divergence

Given an D-dimensional (real-valued) feature vector, $\mathbf{x} \in \mathbb{R}^D$, let $\boldsymbol{X}$ denote the feature space and $\boldsymbol{\mathcal{P}}$ denote the space of probability density functions defined on $\boldsymbol{X}$. For two probability density functions, $P_1, P_2 \in \boldsymbol{\mathcal{P}}$, the Kullback-Leibler (KL) divergence of P_2 from P_1 is defined as [19]:

$$I_{KL}(P_1|P_2) = \int_{\boldsymbol{X}} P_1(\mathbf{x}) \ln\left(\frac{P_1(\mathbf{x})}{P_2(\mathbf{x})}\right) d\mathbf{x} \tag{1}$$

As I_{KL} is an asymmetric divergence measure, i.e., $I_{KL}(P_1|P_2) \neq I_{KL}(P_2|P_1)$, a symmetric KL divergence is defined as [19]:

$$I_{SKL}(P_1, P_2) = \frac{1}{2}|I_{KL}(P_1|P_2) + I_{KL}(P_2|P_1)| \tag{2}$$

Given two GMMs, $\mathcal{G}_1, \mathcal{G}_2 \in \boldsymbol{\mathcal{P}}$, the symmetric KL divergence between them cannot be computed in closed form. Typically an approximation based on MAP adapted GMMs (from a suitable UBM) is utilised [20, 21]. In this work, the GMMs are not obtained via MAP adaptation and hence a Monte-Carlo approximation of the symmetric KL divergence, $\hat{I}_{SKL}$, is used based on

$$\int_{\boldsymbol{X}} f(\mathbf{x})P(\mathbf{x})dx \to \lim_{N\to\infty} \frac{1}{N}\sum_{t=1}^{N} f(\mathbf{x}_t), \qquad as\ N \to \infty \tag{3}$$

where $P \in \boldsymbol{\mathcal{P}}$ and the samples $\mathbf{x}_t$ are assumed to be drawn from $P(\mathbf{x})$. Thus,

$$\hat{I}_{SKL}(\mathcal{G}_1, \mathcal{G}_2) = \frac{1}{2N}\left|\sum_{\mathbf{x}\sim\mathcal{G}_1} \ln \mathcal{G}_1(\mathbf{x}) - \sum_{\mathbf{x}\sim\mathcal{G}_1} \ln \mathcal{G}_2(\mathbf{x}) + \sum_{\mathbf{x}\sim\mathcal{G}_2} \ln \mathcal{G}_2(\mathbf{x}) - \sum_{\mathbf{x}\sim\mathcal{G}_2} \ln \mathcal{G}_1(\mathbf{x})\right| \tag{4}$$

where $\mathbf{x}\sim P$ denotes that $\mathbf{x}$ are i.i.d samples drawn from the probability density function $P(\mathbf{x}) \in \boldsymbol{\mathcal{P}}$ and N is the number of data samples drawn from $\mathcal{G}_1(\mathbf{x})$ and $\mathcal{G}_2(\mathbf{x})$.

4.2 Estimating Variability – KL Divergence

Given a set of GMMs, $\mathbb{G} = \{\mathcal{G}_i : 1 \leq i \leq K\}$, we define the KL model separability, $\Gamma_{KL}(\mathbb{G})$, as the average pairwise KL divergence between all possible pairs of GMMs from the set $\mathbb{G}$. i.e.,

$$\Gamma_{KL}(\mathbb{G}) = \frac{1}{K(K-1)}\sum_{i}\sum_{j,j\neq i} \hat{I}_{SKL}(\mathcal{G}_i, \mathcal{G}_j) \tag{5}$$

It can be seen that a set of GMMs that perform well as a classifier will have a large degree of mutual dissimilarity and consequently a large KL model separability when compared with a set of GMMs that are more similar to each other. (It should be noted that the converse is not true, i.e., a large KL model separability does not necessarily imply the set of GMMs will perform well as a classifier).

From the training dataset (as outlined in section 2), speaker-dependent GMMs, $\mathcal{G}_k^{(j)}$, were trained on data from each speaker j (10 speakers) corresponding to each emotion k (4 emotions). From these GMMs, the speaker specific-emotion model separability scores, γ_e, were estimated from each set of speaker specific emotion models and the emotion-specific speaker model separability scores, γ_s, from each set of emotion specific speaker models.

$$\gamma_e(j) = \Gamma_{KL}\left(\mathbb{G}_e{}^{(j)}\right) \text{ and } \gamma_s(k) = \Gamma_{KL}\left(\mathbb{G}_s^{(k)}\right) \tag{6}$$

where $\mathbb{G}_e{}^{(j)} = \left\{\mathcal{G}_k^{(j)} : \forall k\right\}$ and $\mathbb{G}_s^{(k)} = \left\{\mathcal{G}_k^{(j)} : \forall j\right\}$.

A comparison of γ_e with the speaker-independent emotion model separability score, $\overline{\gamma_e}$, obtained from a set of emotion specific GMMs trained on data from all speakers (i.e., speaker independent models) can be used to estimate the effect of speaker variability on the ability to distinguish between different emotional classes based on the statistical properties of the feature space as modelled by the GMMs. Similarly a comparison of γ_s with the emotion independent speaker model separability score, $\overline{\gamma_s}$, can be used to estimate the effect of emotion variability in the feature space on distinguishing between speakers.

$$\overline{\gamma_e} = \Gamma_{KL}(\mathbb{G}_e) \text{ and } \overline{\gamma_s} = \Gamma_{KL}(\mathbb{G}_s) \tag{7}$$

where $\mathbb{G}_e = \{\mathcal{G}_k : \forall k\}$, $\mathbb{G}_s = \left\{\mathcal{G}^{(j)} : \forall j\right\}$, $\mathcal{G}_k$ is the GMM trained on data from all speakers corresponding to emotion k and $\mathcal{G}^{(j)}$ is the GMM trained on data from speaker j corresponding to all emotions.

The four panels of Fig. 1 compare γ_e and γ_s with $\overline{\gamma_e}$ and $\overline{\gamma_s}$ for two different feature spaces: MFCCs alone and pitch + MFCC + ΔMFCCs (concatenated).

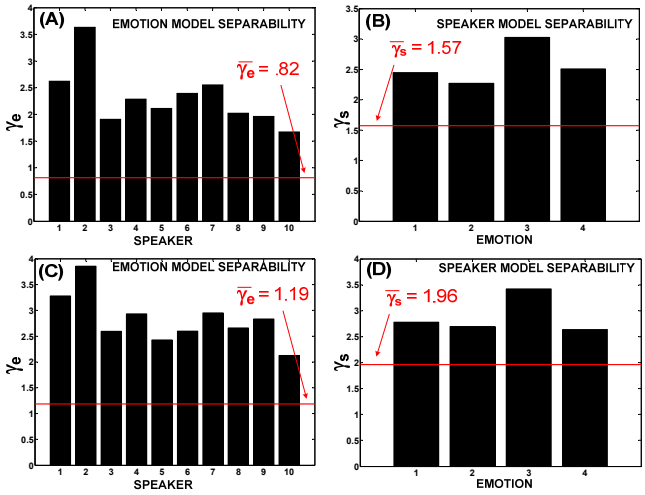

Figure 1: Emotion model separability comparison for (A) MFCC (C) pitch + MFCC + ΔMFCC and speaker model separability comparison for (B) MFCC (D) pitch + MFCC + ΔMFCC. Difference between γ_e and $\overline{\gamma_e}$ estimates the effect of speaker variability on emotion classification and vice versa for γ_s and $\overline{\gamma_s}$

4.3 Estimating Variability – Model Covariance

In addition to quantifying the effect of variability using model separability, an alternative measure may be estimated from the covariance matrices of mixture components of GMMs that model class-conditional probability densities on the feature space. Given a GMM, $\mathcal{G}$, with mixtures constrained to have diagonal covariance matrices, we define the average local variance of the GMM, $\Lambda(\mathcal{G})$, as

$$\Lambda(\mathcal{G}) = \frac{1}{M \cdot D} \mathrm{tr}\left(\hat{\Sigma}^{-1} \sum_{i=1}^{M} \Sigma_i\right) \tag{8}$$

where, $\mathrm{tr}(\cdot)$ denotes the matrix trace, D is the dimensionality of the feature space, M is the number of mixtures in $\mathcal{G}$, Σ_i is the diagonal covariance matrix corresponding to the i-th mixture component and $\hat{\Sigma}$ is the covariance matrix corresponding to a single mixture GMM, $\hat{\mathcal{G}}$, trained on the same data as $\mathcal{G}$ was trained on. $\hat{\Sigma}$ is used to compensate for differences in scale across the different dimensions of the feature space.

Since the different mixture components of a GMM generally take significant (i.e., not almost zero) values on different localised regions of the feature space, the average local variance of a GMM, $\Lambda(\cdot)$, can be thought of as an estimate of the spread of data modelled by the GMM, within clusters in the feature space. This suggests a straightforward way to compare the change in data variability in one model (GMM) compared with another by taking the ratio of their average local variances. Hence, to estimate the effect of speaker variability on emotion models, we estimate the emotion-specific average local variability ratio for each speaker with respect to speaker independent models. We then take the average value across all models as a measure of overall change in localised data spread corresponding to emotion models due to speaker variability, η_e. Since this measure is a ratio, a value greater than one indicates an increase in local spread and vice versa. A similar measure can also be obtained to quantify the change in localised data spread corresponding to speaker models due to emotion variability, η_s.

$$\eta_e = \frac{1}{N_s \cdot N_e} \sum_{k=1}^{N_e} \sum_{j=1}^{N_s} \frac{\Lambda\left(\mathcal{G}_k^{(j)}\right)}{\Lambda(\mathcal{G}_k)} \tag{9}$$

$$\eta_s = \frac{1}{N_s \cdot N_e} \sum_{k=1}^{N_e} \sum_{j=1}^{N_s} \frac{\Lambda\left(\mathcal{G}_k^{(j)}\right)}{\Lambda(\mathcal{G}^{(j)})} \tag{10}$$

where, N_s is the number of speakers, N_e is the number of emotions, $\mathcal{G}_k^{(j)}$ is the GMM trained on data from speaker j corresponding to emotion k, $\mathcal{G}_k$ is the GMM trained on data from all speakers corresponding to emotion k and $\mathcal{G}^{(j)}$ is the GMM trained on data from speaker j corresponding to all emotions.

Table 1 gives the n_e and η_s values estimated from the training dataset (as outlined in section 2) for the two feature spaces: MFCCs and pitch+MFCC+ΔMFCC (concatenated).

Table 1: η_e and η_s values estimated on training set

Feature Space	η_e	η_s
MFCC	0.728	0.801
Pitch + MFCC + ΔMFCC	0.912	0.937

4.4 Speaker Variability in Feature Space Clustering

It is reasonable to assume that the data corresponding to each emotion are distributed in the feature space in clusters (since a lack of any cluster-like structures would suggest there is little or no information contained in the distribution and that the feature is unsuitable for the classification problem at hand). The results reported in Figure 1 and Table 1 lend strong support to the hypothesis that speaker variability affects the distribution of data in the feature space which, in terms of the clusters in the feature space, can mean some combination of shifting of clusters, resizing of clusters and destruction/creation of clusters. If a further assumption is made that the underlying structure of the clusters is representative of the generic acoustic space and that emotion and speaker specific variability manifests as variations to this structure (akin to the assumption made in GMM-UBM based approaches to speaker verification), it is reasonable to expect that most of the variability would manifest as shifting and resizing of clusters.

In order to estimate the relative magnitudes of both effects (shifting and resizing) due to speaker variability on emotion classification, speaker specific-emotion model separability scores, $\widehat{\gamma_e}(j) = \Gamma_{KL}\left(\widehat{\mathbb{G}}_e^{(j)}\right)$ were estimated from speaker specific sets of emotion models with mixture covariances artificially scaled to match the average local variance of the corresponding speaker independent emotion model. Here, $\widehat{\mathbb{G}}_e^{(j)} = \left\{\widehat{\mathcal{G}_k}^{(j)} : \forall k\right\}$ and $\widehat{\mathcal{G}_k}^{(j)}$ is identical to $\mathcal{G}_k^{(j)}$, with the exception that all its covariance matrices are scaled by the factor $\frac{\Lambda(\mathcal{G}_k)}{\Lambda\left(\mathcal{G}_k^{(j)}\right)}$.

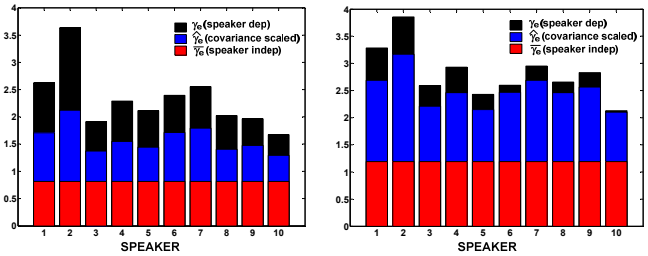

Figure 2: Comparison of γ_e (black), $\widehat{\gamma_e}$ (blue) and $\overline{\gamma_e}$ (red) for: (a) MFCC; (b) pitch+MFCC+ΔMFCC

Comparing γ_e and $\widehat{\gamma_e}$ to $\overline{\gamma_e}$ in Figure 2 suggests that a significant component of the total effect of speaker variability (difference between γ_e and $\overline{\gamma_e}$) is due to shifts in clusters (difference between $\widehat{\gamma_e}$ and $\overline{\gamma_e}$). In particular, for pitch+MFCC+ΔMFCC (Figure 2b), the magnitude of the effect of cluster resizing is small compared to that of cluster shifting.

5. SPEAKER NORMALISATION

The observations made in section 3 suggest that one significant effect of speaker variability on feature vectors is the translation of clusters in the feature space. Hence, one approach to speaker normalisation can be thought of as an attempt to shift these clusters to a location common for all speakers. Joint Factor Analysis (JFA) based channel compensation techniques in speaker verification are designed to exploit similar assumptions regarding speaker and channel variability, and motivate the approach employed here.

Given a M-mixture GMM, $\mathcal{G}$, a supervector representation (taking into account only means) can be defined as $\mathfrak{M}(\mathcal{G}) = [\boldsymbol{\mu}_1^T \ \boldsymbol{\mu}_2^T \ ... \ \boldsymbol{\mu}_M^T]^T$, where $\boldsymbol{\mu}_i \in \mathbb{R}^D$ is the mean of the i-th Gaussian component. The underlying assumption in JFA based normalisation is that $\mathfrak{M}(\mathcal{G})$ can be written as

$$\mathfrak{M}(\mathcal{G}) = \mathfrak{m} + \mathrm{V}\alpha + \mathrm{U}\beta + \mathrm{W}\epsilon \qquad (11)$$

where, $\mathfrak{m} \in \mathbb{R}^{MD}$ is an emotion and speaker independent supervector, $\mathrm{V} \in \mathbb{R}^{MD \times N_V}$ is a matrix of 'eigenemotions' (analogous to eigenvoices), $\mathrm{U} \in \mathbb{R}^{MD \times N_U}$ is a matrix of eigenspeakers (analogous to eigenchannels), $\mathrm{W} \in \mathbb{R}^{MD \times MD}$ is a diagonal matrix, $\alpha \in \mathbb{R}^{N_V}$ represents emotion factors, $\beta \in \mathbb{R}^{N_U}$ represents speaker factors, $\epsilon \in \mathbb{R}^{MD}$ is a random vector and $\mathrm{W}\epsilon$ represents the emotion variability not in the span of the eigenemotions.

In the training phase for the proposed speaker normalisation scheme, a Universal Background Model (UBM), $\mathcal{G}_U$, is estimated from the training set and $\mathfrak{m} = [\overline{\boldsymbol{\mu}}_1^T \ \overline{\boldsymbol{\mu}}_2^T \ ... \ \overline{\boldsymbol{\mu}}_M^T]^T$, where $\overline{\boldsymbol{\mu}}_i$ is the mean of the i-th component of the UBM. From the zeroth and first order Baum-Welch statistics of the training set with respect to the UBM, the hyper-parameters, V, U and W are estimated.

Normalisation is carried out on all feature vectors on a per utterance basis. Let $\mathfrak{U} = \{\mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_T}\}$ be the set of features vectors extracted from all the frames in an utterance. The emotion and speaker factors, α and β, are estimated from the Baum-Welch statistics corresponding to $\mathfrak{U}$ with respect to $\mathcal{G}_U$. Finally, the frame-level normalised feature vectors, $\tilde{\mathbf{x}}_\mathbf{t}$, are computed as:

$$\tilde{\mathbf{x}}_\mathbf{t} = \mathbf{x}_\mathbf{t} - \sum_{i=1}^{M} \varpi_t^{(i)} V_{(i)} \alpha\,, \qquad \forall\, \mathbf{x}_\mathbf{t} \in \mathfrak{U} \qquad (12)$$

where, $\mathbf{x}_\mathbf{t}$ is the raw feature vector, $\mathrm{V}_{(i)} \in \mathbb{R}^{D \times N_V}$ is a sub-matrix of V corresponding to the i-th Gaussian component of $\mathcal{G}$ such that $\mathrm{V} = \left[\mathrm{V}_{(1)}^T \ \mathrm{V}_{(2)}^T \ ... \ \mathrm{V}_{(M)}^T\right]^T$ and $\varpi_t^{(i)}$ is the Gaussian posterior probability of $\mathbf{x}_\mathbf{t}$ corresponding to the i-th mixture of $\mathcal{G}_U$.

While the training phase of the proposed speaker normalisation technique is identical to the estimation of JFA hyper-parameters in speaker verification systems, the normalisation phase differs. Specifically, in the proposed technique the model (supervector) domain normalisation is mapped back to the feature space, allowing for any machine learning paradigm to be applied on the normalised feature space. This mapping process is similar to the mapping process of the feature domain Weiner nuisance modelling [22]. It also allows for other feature domain normalisation techniques to be applied, both before and after the proposed technique if desired. Additionally, mapping to the feature level also means any back-end that operates on frame based features or their derivatives/functionals may be employed in the back-end.

6. EXPERIMENTAL RESULTS

Preliminary emotion classification experiments were carried out with a standard GMM back-end (256 mixture components) to validate the proposed speaker normalisation technique. Only voiced frames were used in both training and testing phases, with voicing being determined by the pitch extraction algorithm [18]. All accuracies reported in this section are unweighted average recall (UAR) of the four emotional classes (cf. section 2 for classes).

The proposed technique has, at the highest level, three controllable parameters: the number of eigenemotions, N_V, and the number of eigenspeakers, N_U and the number of mixtures in the UBM, M. For all the feature spaces on which classification results are reported, N_V and N_U were varied between 2 and 12, taking even numbered values, and M was varied among 64, 128 and 256; from these the highest accuracies are reported in Table 2. From these results it can be seen that the proposed speaker normalisation technique improves the performance of a GMM-based emotion classification systems on all the features vectors that were tested.

Table 2: Unweighted average recall (UAR) for different front-ends with and without the proposed JFA based normalisation.

Feature Space	UAR (%)	
	Without Norm.	With Norm.
Pitch + Energy (M=64, N_V=8, N_U=2)	40.4 %	**40.5 %**
MFCC (M=128, N_V=6, N_U=6)	53.0 %	**54.3 %**
Pitch + MFCC (M=128, N_V=8, N_U=12)	52.8 %	**54.6 %**
MFCC + ΔMFCC (M=64, N_V=10, N_U=2)	54.4 %	**55.3 %**
Pitch + MFCC + ΔMFCC (M=128, N_V=4, N_U=4)	53.3 %	**55.3 %**

7. CONCLUSIONS

This paper has presented a novel analysis of the effect of speaker variability on emotion specific feature vector distributions. The results of the analysis suggest that a significant component of the effects manifests as shifts in clusters of feature vectors. Reversing these shifts can therefore serve as speaker normalisation and the idea forms the core of the proposed JFA based technique. Joint factor analysis in a GMM supervector space provides a framework for modelling translations of clusters in the feature space from an initial model (UBM). Parameters of this framework (JFA hyper-parameters) can be estimated from training data to distinguish translations due to speaker variability from translations due to emotion variability. It is proposed that this framework then be applied to models of any utterance to compensate for any estimated cluster translations due to speaker variability. Furthermore, this model domain compensation is mapped back to the feature domain so that the JFA framework does not place any constraints on any other component of the emotion classification system. Experimental results included in the paper suggest that the proposed technique consistently improves classification performance.

8. ACKNOWLDGEMENT

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

8. REFERENCES

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

[2] Busso, C., Bulut, M., and Narayanan, S. S., "Toward effective automatic recognition systems of emotion in speech," J. Gratch and S. Marsella, Eds., ed: Oxford University Press, 2012.

[3] Schuller, B., Batliner, A., Steidl, S., and Seppi, D., "Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge," *Speech Communication,* **53**, (2011).

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

[5] Sethu, V., Ambikairajah, E., and Epps, J., "Group Delay Features for Emotion Detection," in *INTERSPEECH-2007*, 2007.

[6] Ni, D., Sethu, V., Epps, J., and Ambikairajah, E., "Speaker variability in emotion recognition - an adaptation based approach," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, 2012.

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

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

[9] Busso, C., Metallinou, A., and Narayanan, S. S., "Iterative feature normalization for emotional speech detection," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, 2011.

[10] Schuller, B., Wimmer, M., Arsic, D., Moosmayr, T., and Rigoll, G., "Detection of security related affect and behaviour in passenger transport," in *Interspeech*, Brisbane, 2008.

[11] Ming, L., Metallinou, A., Bone, D., and Narayanan, S., "Speaker states recognition using latent factor analysis based Eigenchannel factor vector modeling," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, 2012.

[12] Rahman, T. and Busso, C., "A personalized emotion recognition system using an unsupervised feature adaptation scheme," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, 2012.

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

[14] Kenny, P., Boulianne, G., Ouellet, P., and Dumouchel, P., "Joint Factor Analysis Versus Eigenchannels in Speaker Recognition," *Audio, Speech, and Language Processing, IEEE Transactions on,* **15**, (2007).

[15] Kenny, P., Boulianne, G., Ouellet, P., and Dumouchel, P., "Speaker and Session Variability in GMM-Based Speaker Verification," *Audio, Speech, and Language Processing, IEEE Transactions on,* **15**, (2007).

[16] Kenny, P., Ouellet, P., Dehak, N., Gupta, V., and Dumouchel, P., "A Study of Interspeaker Variability in Speaker Verification," *Audio, Speech, and Language Processing, IEEE Transactions on,* **16**, (2008).

[17] Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J., Lee, S., and Narayanan, S., "IEMOCAP: interactive emotional dyadic motion capture database," *Language Resources and Evaluation,* **42**, 2008/12/01 (2008).

[18] Talkin, D., "A robust algorithm for pitch tracking (RAPT)," in *Speech Coding and Synthesis*, W. Kleijn and K. Paliwal, Eds., ed New York: Elsevier, 1995, pp. 495-518.

[19] Kullback, S., "Information theory and statistics," *New York: Dover, 1968, 2nd ed.,* **1**, (1968).

[20] Campbell, W. M. and Karam, Z. N., "Simple and efficient speaker comparison using approximate KL divergence," in *Interspeech*, 2010.

[21] Campbell, W. M., Sturim, D. E., Reynolds, D. A., and Solomonoff, A., "SVM Based Speaker Verification using a GMM Supervector Kernel and NAP Variability Compensation," in *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, 2006.

[22] Sturim, D., Torres-Carrasquillo, P., Quatieri, T. F., Malyska, N., and McCree, A., "Automatic Detection of Depression in Speech using Gaussian Mixture Modeling with Factor Analysis," in *INTERSPEECH-2011, Twelfth Annual Conference of the International Speech Communication Association*, 2011.