GMM based Speaker Variability Compensated System for Interspeech 2013 ComParE Emotion Challenge

Vidhyasaharan Sethu[1], Julien Epps[1,2], Eliathamby Ambikairajah[1,2] and Haizhou Li[3,1]

[1]The School of Electrical Engineering and Telecommunications,
The University of New South Wales, Sydney NSW 2052, Australia
[2]National ICT Australia (NICTA), Australian Technology Park, Eveleigh 1430, Australia
[3]Human Language Technology, Institute for Infocomm Research (I²R), Singapore, 138632
v.sethu@unsw.edu.au, j.epps@unsw.edu.au, ambi@ee.unsw.edu.au, hli@i2r.a-star.edu.sg

Abstract

This paper describes the University of New South Wales system for the Interspeech 2013 ComParE emotion sub-challenge. The primary aim of the submission is to explore the performance of model based variability compensation techniques applied to emotion classification and as a consequence of being a part of a challenge, to enable a comparison of these methods to alternative approaches. In keeping with this focused aim, a simple frame based front-end of MFCC and ΔMFCC is utilised. The systems outlined in this paper consists of a joint factor analysis based system and one based on a library of speaker-specific emotion models along with a basic GMM based system. The best combined system has an accuracy (UAR) of 47.8% as evaluated on the challenge development set and 35.7% as evaluated on the test set.

Index Terms: ComParE emotion challenge, emotion classification, speaker normalisation, joint factor analysis.

10.21437/Interspeech.2013-68

1. Introduction

Emotions are expressed via speech through numerous cues, ranging from low-level acoustic cues to high-level linguistic content. Several approaches to speech-based automatic emotion recognition, each taking advantage of a few of these cues, have been explored, e.g. [1-9]. Ideally, the statistical properties of feature vector distributions would vary significantly between different emotions (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.

Typical features used in automatic emotion recognition systems tend to be those based on cepstral coefficients, spectral energy distribution, pitch and loudness. Although these features are extracted on a frame-by-frame basis, the most commonly adopted approach is to estimate statistical parameters (and other functionals) from feature values corresponding to all the frames in an utterance (turn) that is being evaluated. The baseline features for the challenge are a case in point [10]. Given that feature extraction processes do not add information, it is reasonable to hypothesise that the comparatively superior performance of this turn-based approach over a frame-based one is because it reduces the effect of speaker, phonetic and other sources of variability unrelated to emotions. However, the turn-based front-end is not the only approach to reducing these sources of variability and alternatives include techniques that modify the feature vectors directly [11-13] and those that modify the emotion (class) models to either compensate for [14, 15] or adapt to [16, 17] these variations. This emotion sub-challenge provides an opportunity to indirectly test this hypothesis by comparing the relative performances of a number of different approaches adopted by different systems benchmarked on a common test database. In this vein, all the systems described in this paper will employ one or more back-end based approaches to dealing with variability and a common frame based front-end which extracts MFCCs and delta MFCCs.

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 [18]. Moreover, a rich variety of GMM-based model training, adaptation and compensation techniques exist and have been widely used in speaker verification systems. Therefore GMM-based back-ends were employed in all the systems reported in this paper.

2. System Description

2.1. Basic GMM sub-system

The basic system consists of a MFCC+ΔMFCC (12+12 dimensions excluding C_0) front-end computed with 20ms frames with 10ms overlap using a Hamming window. The back-end is based on Gaussian mixture models (GMMs) with a GMM, $\mathcal{G}_k$, trained for each class (k) via ML (maximum likelihood) estimation. The basic system (Figure 1), abbreviated as B herein, does not make use of any normalisation or adaptation techniques. For each utterance, $\mathfrak{U}$, in addition to estimating emotional class ($\bar{k}_B$), the sub-system also computes a measure of confidence of the decision ($\lambda^{(\mathfrak{U})}_{B-system}$), analogous to log-likelihood ratio, as given by eqn (3).

$$\Lambda_K(\mathfrak{U}) = \sum_{t=1}^{T} \log P(\mathbf{x}_t|\mathcal{G}_k) \quad (1)$$

$$\bar{k}_B(\mathfrak{U}) = \arg\max_k \Lambda_k(\mathfrak{U}) \quad (2)$$

$$\lambda^{(\mathfrak{U})}_{B-system} = \max_k \Lambda_k(\mathfrak{U}) - \max_{k,k\neq\bar{k}_B} \Lambda_k(\mathfrak{U}) \quad (3)$$

Where, $\mathbf{x}_t$ is the feature vector corresponding to the t^{th} frame of the utterance $\mathfrak{U}$, $P(\cdot|\cdot)$ denotes conditional probability and $\mathcal{G}_k$ is the GMM trained on the data corresponding to emotion k.

All other systems outlined in this paper are based on this basic system, incorporating one or more refinements. This provides the best basis for the exploration of model-based variability compensation techniques and is in accordance with the previously mentioned aim of this submission to provide a means to compare such approaches to alternative ones that may be adopted by other entries to the challenge. Also, model-based variability compensation techniques play a significant role in most state of the art speaker verification systems and

have been the focus of most of the recent research in that field. Given the extent to which emotion recognition systems have borrowed techniques from speaker verification systems [14, 19-21], a detailed investigation of model-based variability compensation methods is a logical course of action.

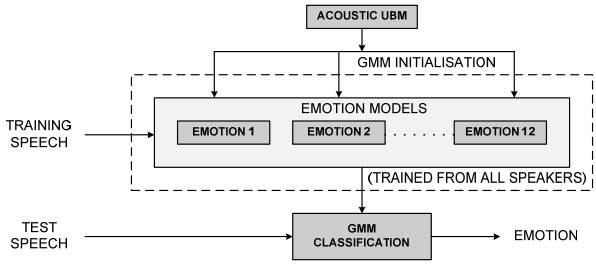

Figure 1: Block diagram of basic GMM sub-system

2.2. Acoustic UBM Seeding

In preliminary experiments, it was observed that the performance of this basic system was somewhat sensitive to the initial conditions (seeds) chosen for the EM algorithm employed in GMM training. Figure 2 demonstrates this by showing the histogram of 12-class emotion classification accuracies in terms of unweight average recalls (UARs) obtained from 100 trials of the basic system (utilising 8-mixture GMMs trained using 10 iterations of the EM algorithm) differing only in the initial seeds for GMM training. Specifically, the initial mixture weights were all initialised identically, the initial mixture covariance vectors (assuming diagonal covariance) were set equal to the training data covariance and the initial mixture means were random vectors drawn from a Gaussian distribution with mean and covariance equal to the training data mean and covariance.

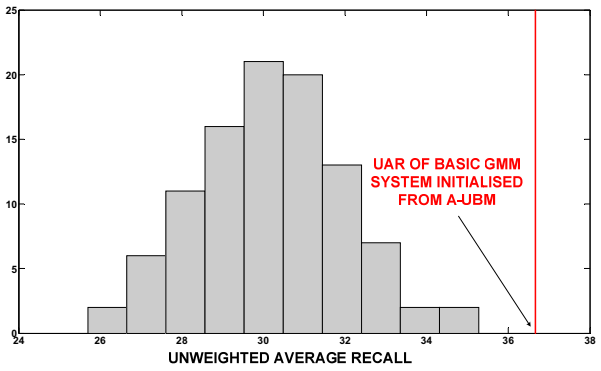

Figure 2: Histogram of 12-class emotion recognition UAR from 100 trials; randomly seeded GMM (8-mixture) systems

To counter this sensitivity, a 1024 mixture GMM was trained on a combination of speech databases to serve as an acoustic universal background model (A-UBM) of the feature space. The databases used to train the A-UBM are the WSJ, WSJCAM0 [22], TIMIT [23], IEMOCAP [24] and AMI [25] corpora. These databases were chosen so as to obtain a background model that spanned a large acoustic space, incorporated a diverse phonetic content and a multitude of speakers and channel conditions. Also, all of these databases consist of speech sampled at 16 kHz, matching the challenge database. Approximately 450 hours (310, 30, 5.5, 12.5 and 90 hours respectively from the above mentioned databases) of speech data was used and apart from IEMOCAP, none of the other databases are 'emotional' speech corpora. To train an N-mixture GMM, with $N < 1024$, the weights, means and covariance vectors corresponding to the 'best' N mixture components of the A-UBM were used as the initial seeds for the EM algorithm. The 'best' mixtures were selected based on occupation counts of the emotion recognition training data, treating the mixtures as hard clusters. The weights were renormalized to make their sum equal to one. Specifically, given a set of training data (for emotion models), $\boldsymbol{\mathcal{X}} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_L\}$, cluster membership estimates of every feature vector were used to determine mixture occupancy counts.

$$\omega_i(\mathbf{x}_t) = \mathcal{N}\left(\mathbf{x}_t; \mu_i^{(A)}, \Sigma_i^{(A)}\right) \tag{4}$$

And,

$$I_\omega(j,t) = \begin{cases} 1, & j = \arg\max_i \omega_i(\mathbf{x}_t) \\ 0, & j \neq \arg\max_i \omega_i(\mathbf{x}_t) \end{cases} \tag{5}$$

$$\mathcal{O}(j) = \sum_{t=1}^{L} I_\omega(j,t), \qquad 1 \leq j \leq 1024 \tag{6}$$

Where, $\mu_i^{(A)}$, $\Sigma_i^{(A)}$ and $w_i^{(A)}$ are the mean, covariance vector and weight corresponding to the i^{th} mixture of the A-UBM; $\omega_i(\mathbf{x}_t)$ denotes the estimate of cluster membership of $\mathbf{x}_t$ towards the i^{th} mixture; $I_\omega(j,t)$ is a binary valued indicator function that takes the value 1 when the j^{th} cluster membership, $\omega_j(\mathbf{x}_t)$, is greater than all the other cluster memberships of $\mathbf{x_t}$ and the value 0 otherwise; and $\mathcal{O}(j)$ is the occupancy count of the j^{th} mixture of the A-UBM.

The occupancy count then forms the basis of selecting N 'best' mixture components of the A-UBM. Let $\mathfrak{N}$ denote this set of N 'best' mixture components. i.e., $\mathfrak{N} \subset \{1, 2, 3, \ldots, 1024\}$ such that $|\mathfrak{N}| = N$ and

$$\mathcal{O}(p) > \mathcal{O}(q), \qquad \forall p \in \mathfrak{N} \; and \; q \notin \mathfrak{N} \tag{7}$$

Where, $|\cdot|$ denotes the number of elements of a set.

The set of means ($\boldsymbol{\mu}$), covariance vectors (Σ) and weights that serve as the seeds for EM training of a N-mixture GMM on $\boldsymbol{\mathcal{X}}$ are then given by:

$$\boldsymbol{\mu} = \left\{\mu_i^{(A)}\right\}, \qquad i \in \mathfrak{N} \tag{8}$$

$$\boldsymbol{\Sigma} = \left\{\Sigma_i^{(A)}\right\}, \qquad i \in \mathfrak{N} \tag{9}$$

$$\mathbf{w} = \left\{\frac{w_i^{(A)}}{\sum_i w_i^{(A)}}\right\}, \qquad i \in \mathfrak{N} \tag{10}$$

Unless otherwise mentioned, all GMMs used in all the systems described in this paper were initialised with the A-UBM. Figure 2 shows the performance of the A-UBM initialised basic system. While this method of seeding the EM algorithm for GMM training is not a speaker variability compensation technique, it eliminated variability in development results due to sensitivity towards initial conditions and thus simplified the system parameter tuning process.

2.3. Speaker-Emotion model library sub-system

Previous work has shown that speaker-specific emotion models are more separable than models trained on data from multiple speakers [21]. It has also been suggested that emotion classification systems can be personalised towards specific speakers and consequently improve their performance by picking speaker-specific emotion models that are close to the target speaker from a set (library) of speaker-specific emotion models following by MAP adaptation with a small amount of development data [16].

For the speaker-emotion model library sub-system (Figure 3), development data from target speakers were not available

for adaptation. However, the core idea of using a library of speaker-specific emotion models (trained from multiple speakers) and picking the best matching models (one per emotion from the library of all speaker-specific models) based on the test utterance for classification was retained, abbreviated as *S* herein. It should be noted that the best matching emotion models need not all correspond to the same speaker. The training data were segregated into speaker-specific datasets and a separate GMM, $\mathcal{G}_k^{(j)}$, was trained for each of the 12 emotional classes (k) corresponding to each of the 10 speakers (j) in the training dataset. For a set of frame based features extracted from a given test utterance, $\mathfrak{U} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_T\}$, the emotional class, $\bar{k}_S$, was estimated as given by eqn (11) and a measure of the confidence of this decision, $\lambda_{S-system}^{(\mathfrak{U})}$ was estimated as given by eqn (14).

$$\bar{k}_S(\mathfrak{U}) = \arg\max_k \widehat{\Lambda}_k(\mathfrak{U}) \tag{11}$$

$$\widehat{\Lambda}_k(\mathfrak{U}) = \max_j \Lambda_k^{(j)}(\mathfrak{U}) \tag{12}$$

$$\Lambda_k^{(j)}(\mathfrak{U}) = \sum_{t=1}^{T} \log P\left(\mathbf{x}_t \middle| \mathcal{G}_k^{(j)}\right) \tag{13}$$

$$\lambda_{S-system}^{(\mathfrak{U})} = \max_k \widehat{\Lambda}_k(\mathfrak{U}) - \max_{k,k\neq\bar{k}_S} \widehat{\Lambda}_k(\mathfrak{U}) \tag{14}$$

Where $\mathbf{x_t}$ is the feature vector corresponding to the t^{th} frame of the utterance $\mathfrak{U}$, $P(\cdot|\cdot)$ denotes conditional probability and $\mathcal{G}_k^{(j)}$ is the GMM trained on the data from speaker *j* corresponding to emotion k.

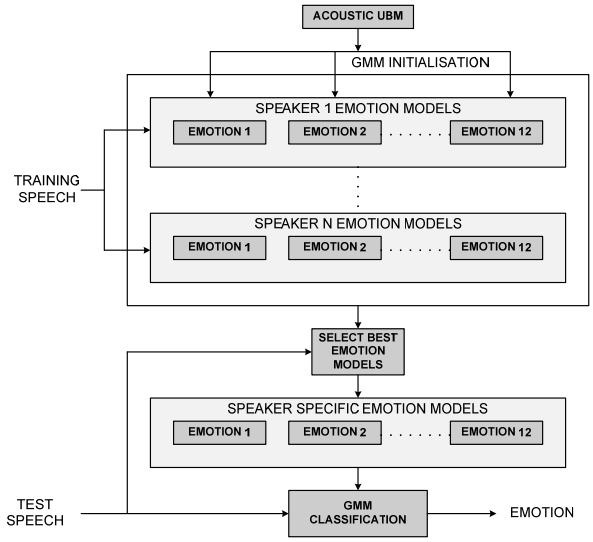

Figure 3: Block diagram of speaker model sub-system

2.4. Joint Factor Analysis based sub-system

The joint factor analysis (JFA) based sub-system adopted a normalisation based approach to dealing with speaker variability, abbreviated as *J* herein. In particular, it utilised a JFA based compensation technique that has been shown to improve the performance of emotion classification systems [21]. Given a M-mixture GMM, $\mathcal{G}$, a supervector representation (taking into account only means) can be defined as $\mathfrak{M}(\mathcal{G}) = [\boldsymbol{\mu}_1^T \ \boldsymbol{\mu}_2^T \dots \ \boldsymbol{\mu}_M^T]^T$, 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

$$\mathfrak{M}(\mathcal{G}) = \mathfrak{m} + \mathrm{V}\alpha + \mathrm{U}\beta + \mathrm{W}\epsilon \tag{15}$$

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 'eigen-emotions' (analogous to eigenvoices), $\mathrm{U} \in \mathbb{R}^{MD\times N_U}$ is a matrix of eigen-speakers (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 eigen-emotions.

At the system training stage, a background GMM, $\mathcal{G}_U$, is estimated from training data from all speakers corresponding to all emotions and $\mathfrak{m} = [\bar{\boldsymbol{\mu}}_1^T \ \bar{\boldsymbol{\mu}}_2^T \dots \ \bar{\boldsymbol{\mu}}_M^T]^T$, where $\bar{\boldsymbol{\mu}}_i$ 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 $\mathcal{G}_U$, the hyper-parameters, V, U and W are estimated.

Normalisation is carried out on all feature vectors on a per utterance basis. Given a set of feature vectors, $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_T\}$, extracted from all the frames in an utterance, $\mathfrak{U}$, 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, $\tilde{\mathbf{x}}_t$, are computed as [21]:

$$\tilde{\mathbf{x}}_t = \mathbf{x}_t - \sum_{i=1}^{M} \varpi_t^{(i)} \mathrm{V}_{(i)}\alpha\,, \qquad \forall\, \mathbf{x}_t \in \mathfrak{U} \tag{16}$$

where, $\mathbf{x}_t$ 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 $\mathrm{V} = \left[\mathrm{V}_{(1)}^T \ \mathrm{V}_{(2)}^T \dots \mathrm{V}_{(M)}^T\right]^T$ and $\varpi_t^{(i)}$ is the Gaussian posterior probability of $\mathbf{x}_t$ corresponding to the i-th mixture of $\mathcal{G}_U$.

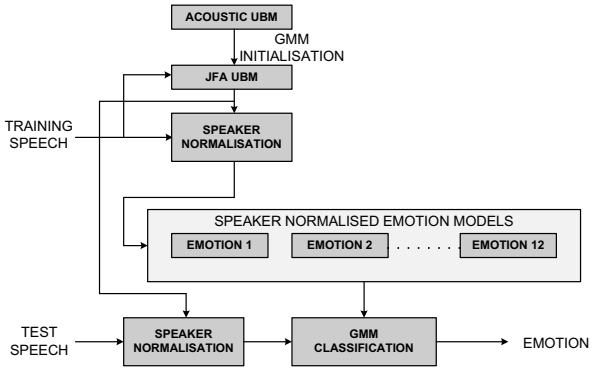

Figure 4: Block diagram of JFA based speaker normalised sub-system

Since the emotion models (GMMs) of this sub-system are trained on the normalised data, they cannot be seeded from the A-UBM. Therefore, initial values of mixture means for the EM algorithm were chosen using the k-means++ algorithm [26] on the class specific training data. Further, since a degree of randomness is inherent in seeding the k-means++ algorithm itself, 100 instances of the JFA sub-system were trained and evaluated on the development set and the best performing one was picked for use. A measure of confidence of each classification decision made by the JFA sub-system, $\lambda_{J-system}^{(\mathfrak{U})}$, was estimated as given in eqn (19).

$$\widetilde{\Lambda}_k(\mathfrak{U}) = \sum_{t=1}^{T} \log P\left(\tilde{\mathbf{x}}_t \middle| \tilde{\mathcal{G}}_k\right) \tag{17}$$

$$\bar{k}_J(\mathfrak{U}) = \arg\max_k \widetilde{\Lambda}_k(\mathfrak{U}) \tag{18}$$

$$\lambda_{J-system}^{(\mathfrak{U})} = \max_k \widetilde{\Lambda}_k(\mathfrak{U}) - \max_{k,k\neq\bar{k}_J} \widetilde{\Lambda}_k(\mathfrak{U}) \tag{19}$$

Where, $\tilde{\mathbf{x}}_t$ is the normalised feature vector corresponding to t^{th} frame of the utterance $\mathcal{U}$ and $\tilde{G}_k$ is the GMM trained on data normalised according to (16), corresponding to emotion k.

3. Sub-System Fusion

The three individual sub-systems are sufficiently distinct to expect a better performance when fused. Two methods for sub-system fusion were employed: selective fusion and linear fusion.

3.1. Selective fusion

In the selective fusion method, for each utterance ($\mathcal{U}$), the decision made by the sub-system with the highest normalised confidence measure, $\bar{\lambda}_i(\mathcal{U})$, was chosen as the final decision. Specifically, the final emotional class decision, k', is determined as follows:

$$k'(\mathcal{U}) = \bar{k}_m(\mathcal{U}) \tag{20}$$

$$m = \arg\max_i \bar{\lambda}_i(\mathcal{U}), \qquad i \in \mathcal{S} \tag{21}$$

Where, $\mathcal{S} \subseteq \{B-system,\ S-system,\ J-system\}$ is the set of sub-systems being fused (all three sub-systems or two of the three). If $\mathcal{D}$ is the set of all utterances in the challenge development set, the normalised confidence measure is given by:

$$\bar{\lambda}_i(\mathcal{U}) = \frac{\lambda_i(\mathcal{U})}{\sum_{X\in\mathcal{D}} \lambda_i(X)} \tag{22}$$

3.2. Linear fusion

The linear fusion system accepts the class posterior probabilities per utterance from each of the sub-systems and performs linear score fusion with sub-system specific offsets. The linear fusion system was implemented using the FoCal multiclass toolkit [27]. The fusion weights were estimated on the challenge development dataset and used to estimate the classification accuracies on both development and test sets.

4. Experimental Results

Parameters of all three sub-systems, the basic GMM sub-system (Section 2.1), the speaker model sub-system (Section 2.3) and the JFA sub-system (Section 2.4) were optimised on the development dataset of the challenge database [10]. Specifically, for the basic GMM sub-system and the speaker model sub-system, the number of mixtures and the number of EM training iterations were chosen based on performance on the development set. For the JFA sub-system, the number of mixtures and the number of training iterations were set to be the same as the basic GMM system and number of eigen-emotions (N_V) and the number of eigen-speakers (N_U) were optimised. During this parameter tuning process it was observed that the system performance on the development set was somewhat sensitive to the parameter values for all three sub-systems (B-,S- and J-), in a manner similar to the previously mentioned sensitivity to initial conditions for model training (Figure 2), exhibiting changes in UAR of around 4% to 5% with small changes in parameter values.

The performances of the individual sub-systems on the development set are listed in Table 2 along with the baseline system UAR [10]. Five of these system configurations, including the 3 individual sub-systems on their own, were chosen for evaluation on the test set and the accuracies obtained in terms of UAR are listed in Table 3. Table 3 also lists the UARs evaluated for arousal and valence with the 12-class labels mapped to the two binary tasks.

Table 1: Sub-system parameters optimised on the challenge development dataset

	Sub-System		
	B-Sys	**S-Sys**	**J-Sys**
No. of Mixtures	128	32	128
No. of EM iteration	10	10	10
No. of eigen-emotions	-	-	10
No. of eigen-speakers	-	-	14

Table 2: 12-class unweighted average recall (UAR) on the challenge development set

System	UAR
Baseline [10]	40.1 %
Basic GMM sub-system (B-sys)	41.0 %
Speaker model sub-system (S-sys)	41.5 %
JFA sub-system (J-sys)	43.1 %
B-sys + S-sys (selective)	38.8 %
B-sys + J-sys (selective)	42.1 %
S-sys + J-sys (selective)	43.8 %
B-sys + S-sys + J-sys (selective)	43.3 %
B-sys + S-sys (linear)	45.6 %
B-sys + J-sys (linear)	42.4 %
S-sys + J-sys (linear)	47.8 %
B-sys + S-sys + J-sys (linear)	45.6 %

Table 3: 12-class unweighted average recall (UAR) on the challenge test set

System	UAR		
	12-Class	**Arousal**	**Valence**
Baseline [10]	40.9 %	75.0 %	61.6 %
Basic GMM sub-system	34.2 %	74.2 %	58.4 %
Speaker model sub-system	33.4 %	72.7 %	59.8 %
JFA sub-system	34.8 %	73.4 %	60.3 %
S-sys + J-sys (selective)	34.2%	70.8 %	60.0 %
S-sys + J-sys (linear)	35.7 %	73.9 %	59.5 %

5. Conclusion

This paper describes our submission to the Interspeech 2013 ComParE emotion sub-challenge. The systems were developed with the specific aim of enabling comparisons between model based approaches to other methods for dealing with speaker variability. As can be seen from the results included in section 4, all three individual systems outperform the baseline on the development database and the best fused system outperforms the baseline system by a significant margin. The system accuracies as evaluated on the test set, however, are worse than the baseline system accuracy suggesting that the systems have been somewhat over-trained towards the development set. Since the development set was never used in the system training phases except to evaluate the linear fusion weights, the mismatch is almost certainly a consequence of 'over-tuning' of the system parameters, specifically the number of mixtures, training iteration, eigen-emotions and eigen-speakers, and not over-fitting of the back-end models in the conventional sense. It is also noteworthy that the results suggest that the best system is a combination of the S-system and the J-system, which each take complementary model based approaches to dealing with speaker variability.

6. Acknowledgements

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

7. References

[1] Barra, R., Montero, J. M., Macias-Guarasa, J., D'Haro, L. F., San-Segundo, R., and Cordoba, R., "Prosodic and Segmental Rubrics in Emotion Identification," in *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, 2006, pp. I-I.

[2] Borchert, M. and Dusterhoft, A., "Emotions in speech - experiments with prosody and quality features in speech for use in categorical and dimensional emotion recognition environments," in *Natural Language Processing and Knowledge Engineering, 2005. IEEE NLP-KE '05. Proceedings of 2005 IEEE International Conference on*, 2005, pp. 147-151.

[3] Lugger, M. and Yang, B., "An incremental analysis of different feature groups in speaker independent emotion recognition," in *ICPhS*, 2007.

[4] Pantic, M. and Rothkrantz, L. J. M., "Toward an affect-sensitive multimodal human-computer interaction," *Proceedings of the IEEE,* vol. 91, pp. 1370-1390, 2003.

[5] Ververidis, D. and Kotropoulos, C., "Emotional speech recognition: Resources, features, and methods," *Speech Communication,* vol. 48, pp. 1162-1181, 2006.

[6] Vidrascu, L. and Devillers, L., "Five emotion classes detection in real-world call center data: the use of various types of paralinguistic features," in *Paraling2007*, 2007.

[7] Yacoub, S., Simske, S., Lin, X., and Burns, J., "Recognition of emotions in interactive voice response systems," in *Eighth European conference on speech communication and technology*, 2003, pp. 729-732.

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

[9] El Ayadi, M., Kamel, M. S., and Karray, F., "Survey on speech emotion recognition: Features, classification schemes, and databases," 10.1016/j.patcog.2010.09.020*, Pattern Recognition,* vol. 44, pp. 572-587, 2011.

[10] Schuller, B., Steidl, S., Batliner, A., Vinciarelli, A., Scherer, K., Ringeval, F., Chetouani, M., Weninger, F., Eyben, F., Marchi, E., Mortillaro, M., Salamin, H., Polychroniou, A., Valente, F., and Kim, S., "The INTERSPEECH 2013 Computational Paralinguistics Challenge: Social Signals, Conflict, Emotion, Autism," in *Interspeech*, Lyon, France, 2013.

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

[12] 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, pp. 5692-5695.

[13] 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, pp. 265-268.

[14] 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, pp. 1937-1940.

[15] 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, pp. 5117-5120.

[16] 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, pp. 5101-5104.

[17] 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, pp. 4948-4951.

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

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

[20] Dumouchel, P., Dehak, N., Attabi, Y., Dehak, R., and Boufaden, N., "Cepstral and Long-Term Features for Emotion Recognition," in *INTERSPEECH-2009*, 2009, pp. 344-347.

[21] Sethu, V., Epps, J., and Ambikairajah, E., "Speaker variability in speech based emotion models - Analysis and normalisation," in *ICASSP*, 2013.

[22] Robinson, T., Fransen, J., Pye, D., Foote, J., and Renals, S., "WSJCAMO: a British English speech corpus for large vocabulary continuous speech recognition," in *Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on*, 1995, pp. 81-84 vol.1.

[23] Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallet, D. S., Dahlgren, N. L., and Zue, V., "TIMIT Acoustic-Phonetic Continuous Speech Corpus," ed. Philadelphia: Linguistic Data Consortium, 1993.

[24] 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," 10.1007/s10579-008-9076-6*, Language Resources and Evaluation,* vol. 42, pp. 335-359, 2008/12/01 2008.

[25] McCowan, I., Carletta, J., Kraaij, W., Ashby, S., Bourban, S., Flynn, M., Guillemot, M., Hain, T., Kadlec, J., and Karaiskos, V., "The AMI meeting corpus," in *Proceedings of the 5th International Conference on Methods and Techniques in Behavioral Research*, 2005.

[26] Arthur, D. and Vassilvitskii, S., "k-means++: the advantages of careful seeding," presented at the Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms, New Orleans, Louisiana, 2007.

[27] Brummer, N., "FoCal Multiclass Toolkit," *URL: http://niko. brummer. googlepages. com/focalmulticlass*.