INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Factor Analysis Based Speaker Normalisation for Continuous Emotion Prediction

Ting Dang [1,2], *Vidhyasaharan Sethu* [1], *Eliathamby Ambikairajah* [1,2]

[1] School of Electrical Engineering and Telecommunications, The University of New South Wales, Sydney, Australia
[2] ATP Research Laboratory, National ICT Australia (NICTA), Australia
ting.dang@student.unsw.edu.au, v.sethu@unsw.edu.au, e.ambikairajah@unsw.edu.au

Abstract

Speaker variability has been shown to be a significant confounding factor in speech based emotion classification systems and a number of speaker normalisation techniques have been proposed. However, speaker normalisation in systems that predict continuous multidimensional descriptions of emotion such as arousal and valence has not been explored. This paper investigates the effect of speaker variability in such speech based continuous emotion prediction systems and proposes a factor analysis based speaker normalisation technique. The proposed technique operates directly on the feature space and decomposes it into speaker and emotion specific sub-spaces. The proposed technique is validated on both the USC CreativeIT database and the SEMAINE database and leads to improvements of 8.2% and 11.0% (in terms of correlation coefficient) on the two databases respectively when predicting arousal.

Index Terms: continuous emotion prediction, speaker normalisation, computational paralinguistics, factor analysis, regression, relevance vector machine

1. Introduction

The dimensional approach for labelling emotions has been attracting increasing attention in recent years in the context of speech based emotion prediction. The primary advantage for using the dimensional approach over discrete emotion labels ('Happy', 'Sad', etc.) is that it has been argued that emotions are a continuum and discrete labels, even with a very large number of them, cannot capture a continuum. The most commonly employed continuous dimensional attributes describing emotions are the two dimensions of 'arousal' and 'valence', with a third 'dominance' dimension increasingly being used as well [1] .

From an engineering point of view, continuous emotion prediction is a regression problem, which outputs a continuous value for the different emotion attributes of interest (valence, arousal and dominance). Currently, most continuous emotion prediction systems adopt high dimensional statistical features as the front-end [2] and employ Support Vector Regression, Long Short-Term Memory Recurrent Neural Network, Gaussian Mixture Regression or Relevance Vector Machines(RVM) in the back-end [3-5]. Among these, RVMs have recently shown to be well suited for emotion recognition [5, 6] and is employed in the systems described in this paper.

Current research on continuous emotion prediction based on speech has primarily focused on either improving the back-end, developing novel features or improving feature selection techniques for choosing the most discriminative feature set from a large pool of (generally statistical) features. Ideally, the chosen features capture only information related to emotional state, but in practice all features also capture acoustic variability (including channel effects), speaker variability, phonetic variability, etc. [7]. The variability not related to emotion information leads to less precise models of emotional states, which in turn introduces errors in the prediction. Among these, speaker variability has been shown to be one of the most significant confounding factor in emotion classification systems that recognise categorical emotion labels (such as 'Happy', 'Angry', etc.) [8] and this is expected to be true for continuous emotion prediction systems as well.

Most speech based inference systems address classification problems such as speech recognition, speaker verification, language identification, etc., and compensation methods for speaker variability inspired by channel compensation in these classification problems have almost universally been developed only for categorical emotion classification systems and not for regression systems that predict continuous emotional attributes. Some of the recently proposed approaches to this speaker variability compensation include: (a) normalisation techniques such as joint factor analysis based normalisation method [9], iterative feature normalisation [10], and an auto-encoder based transfer learning method [11]; (b) model compensation techniques which improves the model representation to decrease the variability [12-14]. Owing to the differences between regression and classification, these normalisation methods (intended for classification problems) cannot be directly applied in a regression framework. Motivated by this lack of speaker normalisation methods for continuous emotion prediction, this paper first investigates the effect of speaker variability and then proposes a normalisation method based on factor analysis to deemphasize the speaker component.

2. Compensation for Speaker Variability

Almost all continuous emotion prediction systems comprise of independent regression systems operating in parallel to predict each attribute of interest such as 'arousal', 'valence', and 'dominance'. Typically each sub-system uses short-term frame based features (low-level descriptors, abbreviated as LLDs) as the basis for prediction and often a larger window spanning multiple frames is used to estimate statistical descriptions of the LLDs corresponding to the frames within this window, and the back-end makes a prediction based on this window level statistical description of LLDs. It is common to employ the

http://dx.doi.org/10.21437/Interspeech.2016-880

same regression method in the back-end and/or the same set of statistical descriptions in the front-end, of all the attribute (arousal, valence, etc.) prediction sub-systems. However, different methods may be used in different sub-systems. The continuous emotion prediction systems used in the experiments described in this paper predict 3 attributes – arousal, valence, and dominance with the sub-systems sharing a common front-end shown in Figure 1.

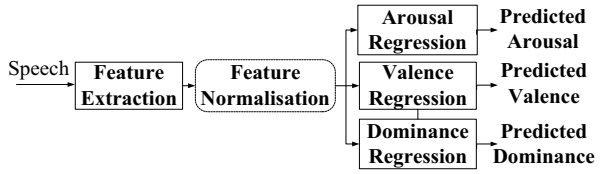

Figure 1: *System Overview – Continuous Emotion Prediction*

2.1. Sub-system description

Detail of the continuous attribute prediction sub-system, employed in all the systems described in this paper, is shown in Figure 2. The Computational Paralinguistics Challenge 2013 (ComParE 2013) audio feature set [15], which contains 65 LLDs and their first derivations, is employed and extracted using OpenSMILE [16]. Five statistical descriptions (functionals), namely, mean, standard deviation, maximum, minimum and range of each feature dimension are applied to the LLDs using a 3s window with a 2s shift between windows. This window size and shift have been previously shown to give good performance [14, 18]. The annotations are averaged within these 3s window as well (to correspond to the statistical features). RVMs are adopted as the back-end since they have previously been shown to perform better than support vector regression, which are more commonly employed [5, 6, 19].

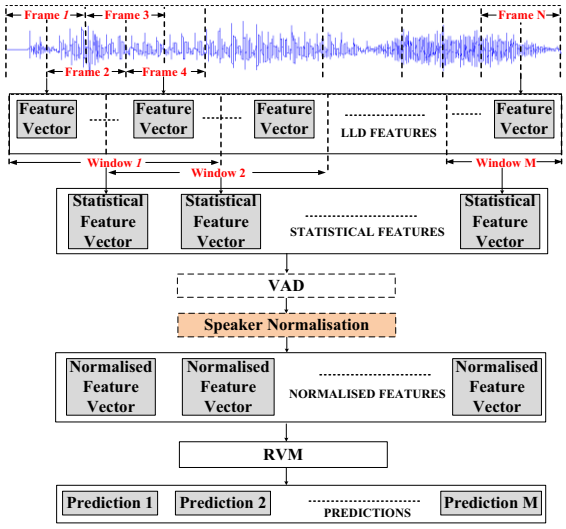

Figure 2: *Sub-system description*

Two types of systems, one with and one without voice activity detection (VAD), are investigated in this paper. One type of system utilises only voiced speech detected by a VAD [17] to train the regression model and unvoiced frames are interpolated after the prediction while the other type uses all frames to train the regression model. The predictions of all systems are smoothed by a binomial filter [6] and the performance of all systems are evaluated in terms of mean correlation coefficient between the predicted attribute values and the ground truth.

2.2. Proposed Speaker Normalisation

The proposed speaker normalisation technique views speaker identity as an underlying factor that affects speech features within a factor analysis framework. Specifically, it assumes features extracted from speech are comprised of a common vector, a speaker identity component and a residual vector that contains mainly emotion-related features (mathematically this is similar to the PLDA model [20]) as given below,

$$x_{ij} = u + \boldsymbol{F}y_i + \varepsilon_{ij} \tag{1}$$

where x_{ij} represents the feature vector estimated from the j^{th} frame of speech from the i^{th} speaker, u is the independent mean over all speakers, y_i is the vector of speaker factors, $\boldsymbol{F}$ is the factor loading matrix that captures the speaker variability, and ε_{ij} is the residual component that contains emotion specific information.

Speaker normalisation is then accomplished by subtracting the speaker identity component, $\boldsymbol{F}y_i$, from the raw features, x_{ij}, to give the normalised features, $\tilde{x}_{ij}$:

$$\tilde{x}_{ij} = x_{ij} - \boldsymbol{F}y_i - u \tag{2}$$

In this model, the speaker factors, y_i, is assumed to follow a standard normal distribution and the residuals, ε_{ij}, is assumed to follow a zero-mean normal distribution with a covariance $\boldsymbol{\Sigma}$. i.e.,

$$y_i \sim \mathcal{N}(0, \boldsymbol{I}) \tag{3}$$

$$\varepsilon_{ij} \sim \mathcal{N}(0, \boldsymbol{\Sigma}) \tag{4}$$

Parameters $\theta = \{u, \boldsymbol{F}, \boldsymbol{\Sigma}\}$ and y_i of the model should be estimated during the training phase using all training speakers' data. The training procedure is identical to that given in [20] .

2.2.1. Model Parameter Estimation

Let $X_i = [{x_{i1}}^T, {x_{i2}}^T \cdots {x_{iM_i}}^T]^T$ represents concatenated window-level features from the i^{th} speaker, where N is the number of speakers and M_i represents the number of feature vectors from the i^{th} speaker. In the training phase, the aim is to find the optimal parameters, θ, that maximises the model likelihood, $P(\boldsymbol{X}|\theta)$, given some training data $\boldsymbol{X}$. Here, the EM algorithm is used to solve the problem as follows:

Equation (1) can be rewritten as:

$$X_i = \begin{bmatrix} u \\ \vdots \\ u \end{bmatrix} + \begin{bmatrix} \boldsymbol{F} \\ \vdots \\ \boldsymbol{F} \end{bmatrix} y_i + \begin{bmatrix} \varepsilon_{i1} \\ \vdots \\ \varepsilon_{iM_i} \end{bmatrix} \tag{5}$$

It is helpful to introduce the notation, $\boldsymbol{A} = [\boldsymbol{F}^T, \boldsymbol{F}^T, \dots, \boldsymbol{F}^T]^T$, $m = [u^T, u^T, \dots, u^T]^T$ and $\varepsilon_i = [\varepsilon_{i1}^T, \varepsilon_{i2}^T, \dots, \varepsilon_{iM_i}^T]^T$ with mean zero and covariance matrix Σ'.

$$\boldsymbol{\Sigma}' = \begin{bmatrix} \boldsymbol{\Sigma} & 0 & \cdots & 0 \\ 0 & \boldsymbol{\Sigma} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \boldsymbol{\Sigma} \end{bmatrix} \tag{6}$$

During the E-step, $P(y_i|X_i, \theta)$ is estimated as:

$$P(y_i|X_i, \theta) = \frac{P(X_i|y_i, \theta)P(y_i)}{P(X_i)} \propto P(X_i|y_i, \theta)P(y_i) \tag{7}$$

Where, the posterior probability $P(X_i|y_i, \theta)$ is a Gaussian distribution as below:

$$P(X_i|y_i, \theta) = \mathcal{N}(m + \boldsymbol{A}y_i, \boldsymbol{\Sigma}') \tag{8}$$

Since the two terms on the right side of Equation (7) are both Gaussian distributions, the posterior distribution is also a Gaussian distribution given as:

$$P(y_i|X_i,\theta) = \mathcal{N}(E[y_i], cov(y_i)) \quad (9)$$

where,

$$E[y_i] = (\boldsymbol{A^T \Sigma'^{-1} A} + \boldsymbol{I})^{-1} \boldsymbol{A^T \Sigma'^{-1}} (X_i - m) \quad (10)$$

$$cov[y_i] = (\boldsymbol{A^T \Sigma'^{-1} A} + \boldsymbol{I})^{-1} \quad (11)$$

In the M-step, the model parameters, $\theta = \{u, \boldsymbol{F}, \boldsymbol{\Sigma}\}$, are optimised to maximise $Q(\theta_{t-1}, \theta_t)$, where t indicates the iteration number.

$$Q(\theta_{t-1}, \theta_t) = \sum_{i=1}^{N} \sum_{j=1}^{M_i} \int P(y_i|X_i,\theta_{t-1}) \log[P(x_{ij}|y_i,\theta_t) P(y_i)] dy_i \quad (12)$$

The updated parameters θ can be obtained by calculating the derivatives of $Q(\theta_{t-1}, \theta_t)$ and are given as:

$$u = \frac{1}{N * M_i} \sum_{i=1}^{N} \sum_{j=1}^{M_i} x_{ij} \quad (13)$$

$$\boldsymbol{F} = (\sum_{i=1}^{N} \sum_{j=1}^{M_i} (x_{ij} - u) E[y_i]^T)(\sum_{i=1}^{N} E[y_i y_i^T])^{-1} \quad (14)$$

$$\boldsymbol{\Sigma} = \frac{1}{N * M_i} \sum_{i=1}^{N} \sum_{j=1}^{M_i} diag[(x_{ij} - u)(x_{ij} - u)^T - \boldsymbol{F} E[y_i](x_{ij} - u)^T] \quad (15)$$

2.2.2. Speaker Normalisation for Test Utterances

During the test phase, the speaker factors, y_t, are estimated from $P(y_t|z_t,\theta)$, where z_t represents the test data. The test speaker factor y_t is estimated as the expected value, $E[y_t]$, given by (16), where $\boldsymbol{A} = \boldsymbol{F}$ and $\boldsymbol{\Sigma'} = \boldsymbol{\Sigma}$ since the normalisation is carried out at the window-level, and the normalised feature vectors, $\tilde{z}_t$, are calculated as given by equation (17).

$$E[y_t] = (\boldsymbol{F^T \Sigma^{-1} F} + \boldsymbol{I})^{-1} \boldsymbol{F^T \Sigma^{-1}} (z_t - u) \quad (16)$$

$$\tilde{z}_t = z_t - \boldsymbol{F} E[y_t] - u \quad (17)$$

3. Database

The USC CreativeIT database [21] and the SEMAINE database [22] were utilised to evaluate the proposed method. The CreativeIT database is an audio-visual database recorded using the theatrical improvisation technique of Active Analysis. Spontaneous dialogues and acted dialogues are both recorded in a dyadic conversation. It contains 8 sessions of 90 sentences recorded from 16 speakers, each of which consists of 6-14 sentences. The annotation contains continuous rating of arousal, valence and dominance attributes obtained by asking raters to watch the video sessions and use FEELTRACE tool [23] that can be continuously moved to record their perceived emotion attribute values (values lie between -1 and 1). The final continuous attribute values were obtained by averaging all individual annotations.

The SEMAINE database uses the Sensitive Artificial Listener paradigm to record natural conversations between a person and an operator (role-played by a person). The operator assumes one of four personalities to elicit different emotion states of the user whose data was used in the experiments. In total, speech data from 18 speakers (users) recorded over 24 sessions are used in the experiment. Annotations are carried out using FEELTRACE tool. The overall arousal, valence and dominance (power) ratings were obtained by averaging the corresponding ratings from 6-8 raters.

4. Experimental Results

Three experiments were conducted to establish the impact of speaker variability on the performance of continuous emotion prediction systems, to investigate the effect of the proposed speaker normalisation technique on the feature space and to validate its use in an emotion prediction system.

4.1. Impact of Speaker Variability

In order to determine if speaker variability had a significant negative impact on the performance of continuous emotion prediction systems, the performance of a speaker independent emotion prediction system was compared to that of speaker-specific emotion prediction systems on the USC CreativeIT database. Speaker-specific emotion prediction systems refer to those that are trained and tested on data from the same speaker. For this experiment, speaker-specific systems were trained on 2/3rd of the data and tested on the remaining 1/3rd of the data from 14 of the 16 speakers in the database (there was insufficient data from the other 2 to train and test a speaker-specific system). The performance of the speaker independent system is estimated on data from all 8 sessions in the database in a leave-one-session-out cross-fold validation. Both systems only use voiced speech for training (refer to section 2.1) and no feature normalisation is employed.

The results of the experiment are shown in Figure 3, where the performance of the 14 speaker-specific system as well as the average speaker-specific performance is compared to the performance of the speaker independent system in terms of mean correlation coefficient between predicted attribute values and ground truth labels based on human annotators (included in the database). The superior performance of speaker-specific system in general suggests that speaker variability degrades the performance of speech based continuous emotion prediction systems. It should be noted that the speaker-independent models are trained with approximately 15 times as much data as the speaker-specific systems and consequently the comparison is intended to be indicative only and not definitive.

4.2. F-ratio measurement

In order to investigate the effect of the normalisation on raw features, F-ratio is a used as a measure of dissimilarity between speaker classes [24]. It is the ratio of inter-class variability over intra-class variability given by:

$$F\_ratio = \frac{\frac{1}{N}\sum_{i=1}^{N}(u_i - u)}{\frac{1}{N \cdot M_i}\sum_{i=1}^{N}\sum_{j=1}^{M_i}(x_{ij} - u_i)} \quad (18)$$

Where u_i represents the mean of features estimated from the i^{th} speaker and other notations are kept same as section 2.2.

In this experiment, we treat each speaker as a distinct class and adopt the average F-ratio of speaker classes as a measure of feature dissimilarity between speakers per feature dimension. A larger F-ratio value indicates a more separated feature and therefore greater speaker variability.

The features employed in the systems outlined in section 2.1 are used in this experiment as well. The proposed speaker

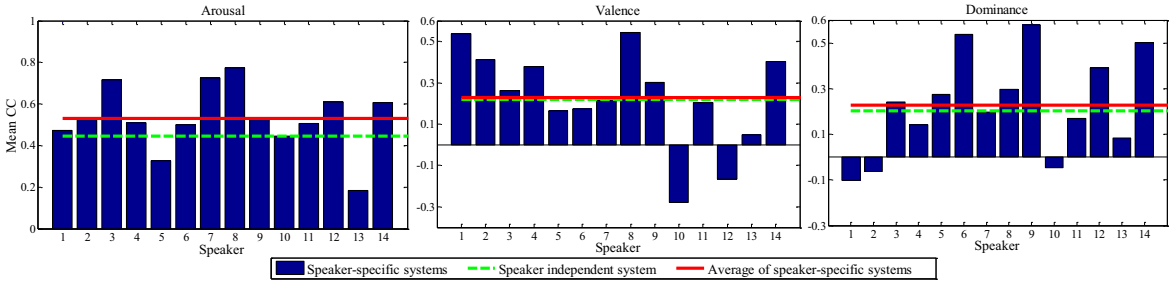

Figure 3: *Speaker independent vs. speaker-specific systems - correlation coefficient evaluated on the USC CreativeIT database.*

normalisation method is applied with 13-dimensional speaker factor vectors (one less than the number of speakers in the training-set in each fold of cross-fold validation) and 10 iterations of EM algorithm for parameter estimation. F-ratios of the first 50 dimensions of the un-normalised feature vector are compared with the F-ratios of corresponding 50 dimensions of the normalised feature vector in Figure 4. Here only the first 50 out of 650 are shown in order to reduce clutter and make the graph readable but the relative relationships between original features, speaker identity component ($\boldsymbol{FE}[y_t]$) and normalised features were observed to be generally consistent across all 650 dimensions. In addition the F-ratios of the same 50 dimensions of the speaker identity component are also shown in Figure 4. From the figure it can be seen that consistently the largest F-ratios correspond to the speaker identity component and the smallest F-ratios to the normalised feature vectors which suggests that the proposed speaker normalisation method is operating as expected and is able to decompose the feature space into a speaker subspace and a residual subspace (which includes emotion information).

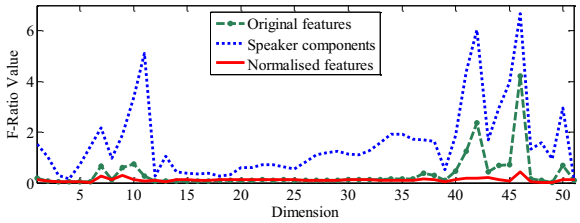

Figure 4: *F-ratio comparison among original features, speaker component and normalised features*

4.3. Factor analysis based speaker normalisation

The final validation of the proposed speaker normalisation technique was carried out on both the USC CreativeIT database and the SEMAINE database by comparing the performances of the basic emotion prediction systems outlined in section 2.1 with and without speaker normalisation.

The proposed technique was applied with 13-dimensional and 12-dimensional speaker factor vectors with the USC CreativeIT and the SEMAINE databases respectively. The dimensionality of the speaker factor vectors were chosen based on the number of speakers in the training dataset. The speaker normalisation model parameters were estimated with 10 iterations of the EM algorithms in both cases.

The experiments on the USC CreativeIT database were carried out in a leave-one-session-out cross validation manner. The SEMAINE database on the other hand was split into a distinct training set comprising of speech data from 12 randomly selected speakers and a distinct test comprising of speech from the remaining 6 speakers. The accuracies of continuous emotion prediction systems with and without the proposed normalisation are shown in Table 1. As mentioned in section 2.1, two versions of each system were tested – one using only voiced speech and another using all speech.

Table 1. *Performance on two databases. **A** means arousal, **V** means valence and **D** means dominance.*

		Mean Correlation Coefficient (CC)		
		A	**V**	**D**
USC CreativeIT Database	Model with VAD	0.447	0.220	0.201
	Model with VAD + Normalisation	***0.483***	***0.246***	***0.215***
	Model without VAD	0.527	0.238	0.237
	Model without VAD +Normalisation	0.526	0.231	0.220
SEMAINE Database	Model with VAD	0.453	0.106	0.623
	Model with VAD + Normalisation	***0.503***	***0.208***	***0.635***
	Model without VAD	0.429	0.116	0.611
	Model without VAD +Normalisation	***0.521***	***0.211***	***0.643***

It can be seen the proposed speaker normalisation consistently improves the performance of emotion prediction systems that use voiced speech on both databases, which is relatively 8.2%, 11.7% and 7% in USC CreativeIT and 11.0%, 95.7% and 1.9% in SEMAINE for arousal, valence and dominance respectively. However, it does not show improvement in the system that uses all frames in USC CreativeIT database. This may be due to the slightly higher proportion of unvoiced speech in the USC CreativeIT database (16% of all frames) when compared with the SEMAINE database (5% of all frames). Finally it should be noted that the only other published system evaluated using audio data only from the USC CreativeIT database using an identical cross-fold validation reported a mean correlation coefficient of 0.478 for arousal and 0.133 for dominance (valence was not reported) [14].

5. Conclusion

This paper investigated the negative impact of speaker variability on continuous emotion prediction system and proposed a novel factor analysis based normalisation method. The normalisation technique was validated on both the CreativeIT and SEMAINE databases and shown to be particularly effective on voiced speech. In addition, analyses of the proposed decomposition of the feature space based on F-ratio of the different components revealed that the technique was able to isolate speaker variability reasonably well. As the first speaker normalisation technique proposed for continuous emotion prediction, it opens up avenues for further improvement.

6. References

[1] H. Gunes, B. Schuller, M. Pantic, and R. Cowie, "Emotion representation, analysis and synthesis in continuous space: A survey," in *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, 2011, pp. 827-834.

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

[3] H. Gunes and B. Schuller, "Categorical and dimensional affect analysis in continuous input: Current trends and future directions," *Image and Vision Computing,* vol. 31, pp. 120-136, 2013.

[4] A. Metallinou, A. Katsamanis, and S. Narayanan, "Tracking continuous emotional trends of participants during affective dyadic interactions using body language and speech information," *Image and Vision Computing,* vol. 31, pp. 137-152, 2013.

[5] M. A. Nicolaou, H. Gunes, and M. Pantic, "Output-associative rvm regression for dimensional and continuous emotion prediction," *Image and Vision Computing,* vol. 30, pp. 186-196, 2012.

[6] Z. Huang, T. Dang, N. Cummins, B. Stasak, P. Le, V. Sethu*, et al.*, "An Investigation of Annotation Delay Compensation and Output-Associative Fusion for Multimodal Continuous Emotion Prediction," in *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*, 2015, pp. 41-48.

[7] C. Busso, M. Bulut, and S. Narayanan, "Toward effective automatic recognition systems of emotion in speech," *Social emotions in nature and artifact: emotions in human and human-computer interaction, J. Gratch and S. Marsella, Eds,* pp. 110-127, 2012.

[8] V. Sethu, E. Ambikairajah, and J. Epps, "Phonetic and speaker variations in automatic emotion classification," in *INTERSPEECH*, 2008, pp. 617-620.

[9] V. Sethu, J. Epps, and E. Ambikairajah, "Speaker variability in speech based emotion models-Analysis and normalisation," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, 2013, pp. 7522-7526.

[10] C. Busso, S. Mariooryad, A. Metallinou, and S. Narayanan, "Iterative feature normalization scheme for automatic emotion detection from speech," *Affective Computing, IEEE Transactions on,* vol. 4, pp. 386-397, 2013.

[11] J. Deng, Z. Zhang, E. Marchi, and B. Schuller, "Sparse autoencoder-based feature transfer learning for speech emotion recognition," in *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*, 2013, pp. 511-516.

[12] P. Lopez-Otero, L. Docio-Fernandez, and C. Garcia-Mateo, "iVectors for Continuous Emotion Recognition," *Training,* vol. 45, p. 50, 2014.

[13] K. W. Gamage, V. Sethu, P. N. Le, and E. Ambikairajah, "An i-vector GPLDA system for speech based emotion recognition," in *2015 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, 2015, pp. 289-292.

[14] H. Khaki and E. Erzin, "Continuous Emotion Tracking Using Total Variability Space," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[15] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval*, et al.*, "The INTERSPEECH 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism," 2013.

[16] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*, 2010, pp. 1459-1462.

[17] M. Brookes, "VOICEBOX: A speech processing toolbox for MATLAB. 2006," ed.

[18] F. Ringeval, F. Eyben, E. Kroupi, A. Yuce, J.-P. Thiran, T. Ebrahimi*, et al.*, "Prediction of asynchronous dimensional emotion ratings from audiovisual and physiological data," *Pattern Recognition Letters,* vol. 66, pp. 22-30, 2015.

[19] M. E. Tipping, "Sparse Bayesian learning and the relevance vector machine," *The journal of machine learning research,* vol. 1, pp. 211-244, 2001.

[20] S. J. Prince and J. H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, 2007, pp. 1-8.

[21] A. Metallinou, C.-C. Lee, C. Busso, S. Carnicke, and S. Narayanan, "The USC CreativeIT database: A multimodal database of theatrical improvisation," *Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality,* p. 55, 2010.

[22] G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schröder, "The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent," *Affective Computing, IEEE Transactions on,* vol. 3, pp. 5-17, 2012.

[23] R. Cowie, E. Douglas-Cowie, S. Savvidou*, E. McMahon, M. Sawey, and M. Schröder, "'FEELTRACE': An instrument for recording perceived emotion in real time," in *ISCA tutorial and research workshop (ITRW) on speech and emotion*, 2000.

[24] X. Lu and J. Dang, "An investigation of dependencies between frequency components and speaker characteristics for text-independent speaker identification," *Speech communication,* vol. 50, pp. 312-322, 2008.