# An i-vector GPLDA System for Speech based Emotion Recognition

Kalani Wataraka Gamage[*†], Vidhyasaharan Sethu[*], Phu Ngoc Le[*†] and Eliathamby Ambikairajah[*†]
[*]School of Electrical Engineering and Telecommunications, UNSW, Australia
[†]ATP Research Laboratory, National ICT Australia (NICTA), Australia
E-mail: kalani.watarakagamage@student.unsw.edu.au

***Abstract*— In this paper, we propose the use of a Gaussian Probabilistic Linear Discriminant Analysis (GPLDA) back-end for utterance level emotion classification based on i-vectors representing the distribution of frame level MFCC features. Experimental results based on the IEMOCAP corpus show that the GPLDA back-end outperforms an SVM based back-end while being less sensitive to i-vector dimensionality, making the proposed framework more robust to parameter tuning during system development.**

## I. INTRODUCTION

The ability to automatically recognize emotions from speech is a highly desirable attribute to have in human-machine interfaces and has potential applications in a number of scenarios ranging from call centers to smart devices. Consequently speech based emotion recognition has been an active area of research and has been so over the past decade [1, 2, 3]. The majority of this research has focused on identifying suitable features, feature selection methods and classification techniques [1, 2, 3]. Currently, most state-of-the-art systems use high dimensional feature sets derived from a number of phonetic, acoustic and prosodic features with an appropriately trained classifier such as Support Vector Machine(SVMs) or Neural Networks as back-ends [2,3]. Some systems optionally use feature selection algorithms to reduce the dimensionality of the initial large feature set prior to classification [4].

An alternative approach is to model the distribution of frame based features such as Mel Frequency Cepstral Coefficients (MFCCs) from each utterance using Gaussian mixture models (GMMs) and then use a vectorial representation of these models (such as supervectors) as the input to the back-end classifiers [5, 6]. In particular, factor analyses of the supervector spaces leading to the i-vector framework, which is the de facto standard in speaker verification systems [7, 8], has shown promise in emotion recognition systems [9, 10] but has not been fully explored.

In [9], an approach based on extended i-vectors was suggested for speech based emotion detection. Use of i-vectors has also been suggested for continuous emotion recognition in [10]. In [5], latent factor analysis approach was shown to improve performance when applied to emotion recognition. Specifically, speaker state factors were extracted from emotional speech and modeled by the back-end. Following i-vector extraction, a number of back-ends have been used as classifiers in speaker verification systems, including cosine distance scoring and support vector machines (SVMs). However, the recently introduced GPLDA based back-ends have emerged as the most effective classification methodology within the i-vector framework for speaker recognition [11]. Our work investigates the suitability of this i-vector GPLDA approach for automatic emotion recognition of categorical emotional labels.

GPLDA is a principled method for manipulating the input features, in this case the i-vectors, such that more discriminative feature subspaces have a greater impact on classification [12]. Consequently, the work reported in this paper also investigates if the GPLDA back-end can handle redundancy in the i-vector representation better than support vector machines.

## II. THE I-VECTOR GPLDA FRAMEWORK

The proposed emotion recognition system (Fig. 1) represents each utterance as an i-vector estimated from frame based MFCCs features. Following this, a GPLDA back-end trained on the i-vectors is used to assign the best matching emotion label to each test utterance.

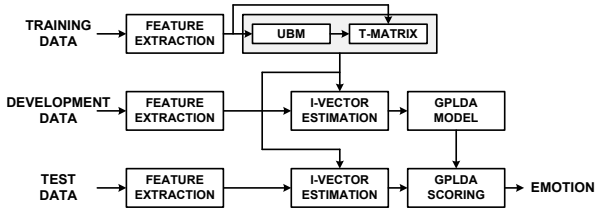


Fig. 1 Overview of proposed i-vector GPLDA System

### A. I-vector Estimation

The GMM supervector represents the probability distribution of short term features extracted from a speech segment, in this case frame-level MFCCs extracted from each utterance [3]. Joint factor analysis (JFA) and the i-vector paradigms aim to overcome limitations of the supervectors (due to high dimensionality) by modelling the variability (information) contained in the supervectors with a much smaller set of factors. Specifically, in the i-vector framework [7], variations in speech due to different factors are modelled by a single space called total variability space while JFA based approaches attempt to model variations due to different factors independently [7]. The total variability space is defined by the total variability matrix **T** comprising of eigenvectors corresponding to largest eigenvalues of the total variability covariance matrix [7]. In the context of emotional speech data, the total variability space represent all the variations in the distributions of the frame level features,

including speaker, emotional and channel variabilities. In this space, a given utterance is modeled as $\mathbf{M} = \boldsymbol{m} + \mathbf{T}\boldsymbol{\eta}$. Where, **M** is the speaker, emotion and channel dependent supervector, $\boldsymbol{m}$ is a speaker, emotion and channel independent supervector, **T** is a low rank rectangular matrix and $\boldsymbol{\eta}$ is the i-vector that represents all the variability in the supervectors in a much lower dimensional vector. Both the total variability matrix **T** and the Universal Background Model (UBM)are estimated from training data comprising of speech corresponding to multiple emotion from multiple speakers.

### B. GPLDA based Back-End

#### 1) GPLDA:

Gaussian probabilistic linear discriminant analysis (GPLDA) is a powerful technique that factorises the input features, in this case the i-vectors, in terms of a small number of underlying identity variables [11, 14]. The assumptions that the probability distribution of the input features given the underlying identity variables is a Gaussian distribution and that the probability distribution of the underlying identity variabilities is also a Gaussian distribution leads to the name 'Gaussian' PLDA. In this work, the i-vectors extracted from each utterance are the input features to the GPLDA back-end. Specifically, given a set of R utterances, the i-vector $\boldsymbol{\eta_r}$ can be modeled as follows:

$$\boldsymbol{\eta_r} = \boldsymbol{\mu} + \boldsymbol{\Phi\beta_r} + \boldsymbol{\varepsilon_r} \quad (1)$$

where r = 1...R, the i-vector model $\boldsymbol{\eta_r}$ is the sum of the emotion specific component, $\boldsymbol{\Phi\beta_r}$ , the noise component, $\boldsymbol{\varepsilon_r}$, and a global offset $\boldsymbol{\mu}$. The matrix $\boldsymbol{\Phi}$ represents the projection from the i-vectors to the underlying latent emotion identity variables represented by $\boldsymbol{\beta}$. The within class variance and residual noise is represented by $\boldsymbol{\varepsilon_r}$ which is characterized by a full covariance matrix $\boldsymbol{\Sigma}$, and $\boldsymbol{\mu}$ is estimated as the mean value of all the training data. The GPLDA model also makes the following assumptions:

$$\boldsymbol{P}(\boldsymbol{\eta_r}|\boldsymbol{\beta_r}, \boldsymbol{\theta}) = \boldsymbol{G}[\boldsymbol{\mu} + \boldsymbol{\Phi\beta_r}, \boldsymbol{\Sigma}] \quad (2)$$

where $G[\boldsymbol{a}, \boldsymbol{b}]$ represents a Gaussian distribution with mean $\boldsymbol{a}$ and covariance matrix $\boldsymbol{b}$. The model parameters $\boldsymbol{\theta}$ = { $\boldsymbol{\mu}$, $\boldsymbol{\Phi}$, $\boldsymbol{\Sigma}$ } are learned from the development data, a separate data set from that used to train the i-vector modeling framework. These model parameters are extracted using EM algorithm as in [11, 14]. It is also assumed that the emotion identity variables have a standard normal distribution, i.e., $P(\boldsymbol{\beta_r}) = G[\mathbf{0}, \mathbf{I}]$.

#### 2) Whitening of i-vectors for GPLDA modeling:

In the GPLDA modeling, two main assumptions are made. They are (i) Identity variable, $\boldsymbol{\beta}$ , and other sources of variabilities (such as speaker, channel and session effects) are independent and; (ii) they have Gaussian probability distributions. But the Gaussianity assumption of the data does not generally hold for speech data [11]. Therefore a normalization method is applied to make the i-vector distribution more Gaussian [11]. Specifically, a linear whitening transform followed by length normalization is used to reduce the non-Gaussianity of the i-vectors [11]. In the systems reported in this paper, the GPLDA models were trained on development data drawn from all emotional classes from multiple speakers. The whitening normalization was carried out for both development and test data.

#### 3) GPLDA Scoring:

For each test utterance, a score is estimated with respect to each i-vector in the development data set [11, 14]. This score is the log likelihood ratio between two competing hypotheses: $H_0$ and $H_1$. Where $H_0$ hypothesizes that both the given test i-vector: $\boldsymbol{\eta_t}$ and considered development i-vector: $\boldsymbol{\eta_d}$ share the same emotion identity variable $\boldsymbol{\beta_1}$ while $H_1$ hypothesizes that they correspond to two distinct emotion identity variables $\boldsymbol{\beta_1}$ and $\boldsymbol{\beta_2}$.

$$\boldsymbol{score} = \boldsymbol{log}\big(\boldsymbol{P}(\boldsymbol{\eta_t}, \boldsymbol{\eta_d}|\, \boldsymbol{H_0})\big) - \boldsymbol{log}(\boldsymbol{P}(\boldsymbol{\eta_t}, \boldsymbol{\eta_d}|\, \boldsymbol{H_1})) \quad (3)$$

In the proposed emotion classification system, in order to associate a specific emotion label to each test utterance, it is scored against all development utterances from each emotional class and the average score per emotion is computed. The test vector is then associated with the emotion corresponding to the highest average score.

## III. EXPERIMENTS

### A. IEMOCAP Database

The Interactive Emotional Dyadic Motion Capture (IEMOCAP) database was used in all the experiments reported in this paper [15]. The database consists nearly 12 hours of audio and visual data collected from 10 actors (5 male and 5 female) engaged in scripted and improvised dialogs in dyadic sessions (1 male and 1 female in each session). The sessions were later manually segmented into utterances and annotated by human evaluators with one or few of the following class labels: angry, happy, sad, neutral, frustrated, excited, fearful, surprised, disgusted, and other.

In the experiments reported in this paper, only the utterances with a majority agreement on the emotion category between evaluators (at least 2 out of 3 evaluators picked the same emotional category) were selected. Further, only the four emotions with maximum number of utterances across the database were considered in the experiments, namely angry, happy, sad and neutral. Finally, in order to balance data distribution among classes, the emotional categories of happy and excited were merged to a one happy class. The data setup is identical to that used in experiments reported in [4, 5, 6, 16, 21].

### B. Experiment Settings

The proposed i-vector GPLDA system was compared to 3 other systems that spans the current approaches to speech based emotion recognitions. Specifically, the following three baseline systems were implemented in addition to the proposed system.

• *Baseline 1: emo-ISO9-SVM:* This system utilizes the 2009Interspeech feature set (emo ISO9 feature set) extracted

using the openSMILE toolkit [17]. These features were then input to a Linear Polynomial Kernel SVM classifier trained using Sequential Minimal Optimization (SMO).

• *Baseline 2: MFCC-Supervector-SVM*: This system uses GMM supervectors [6] estimated from frame based MFCCs as features to the back-end which was identical to the emo-ISO9-SVM above.

• *Baseline 3: MFCC-i-Vector-SVM :* This system modifies the above MFCC-Supervector-SVM system by first estimating i-vectors from the GMM supervectors prior to the SVM based emotion classification identical to the above two systems.

The proposed i-vector GPLDA system and the three baseline systems described above were evaluated using 10-fold leave-one-speaker-out cross validation. For each fold, the development data and training data for a system were obtained by dividing the available data from 9 speakers into two non-overlapping data portions with identical data distributions among all four emotion classes. The held out data from the $10^{th}$ speaker was used as the test data in that fold. Further, since each speaker in the database contain both scripted and improvised speech utterances, care was taken to ensure that equal portions of improvised and acted data from each speaker was assigned to the two non-overlapping training and development datasets in each fold.

In order to implement all the systems, frame level MFCCs were extracted using 50ms windows with 50% overlap between frames. A 39 dimensional feature vector comprising of the first 12 MFCCs, energy, and their first and second derivatives computed per frame. Following feature extraction, the feature vectors corresponding to unvoiced regions were removed using the Voice Activity Detector (VAD) developed in [18]. The i-vector framework and the GPLDA back-end were both implemented using the MSR Identity Toolbox [19], and the SVM back-ends were implemented using WEKA [20]. The number of mixture components in the UBM was varied in order to obtain the best performance on both the baseline and the proposed systems. The effect of i-vector dimensionality on the performance of the proposed approach is reported in section IV (Fig. 2 and Fig. 3).

## IV. RESULTS

In addition to the three baseline systems implemented as part of this work, the results of the proposed i-vector GPLDA system are also compared against a number of other systems previously reported in the literature [4, 5, 16, 21] that have also been tested on the same database, using similar data selection schemes following leave-one-speaker-out cross validation. Table I compares the performance of the proposed system with that of three baseline systems (refer section III B) as well as six systems previously reported in the literature. It should be noted that for these six systems, the classification accuracy results reported by the authors in the literature are repeated here. All the systems share same database and tested under identical conditions. Both Unweighted Average Recall (UAR) and Weighted Average Recall (WAR) are presented, but it should be noted that with unbalanced data sets, UAR may be a better indicator of system performance compared to WAR. The unweighted average recall (UAR) is also the performance metric used in both INTERSPEECH emotion recognition challenges to date.

TABLE I
SUMMARY OF THE CLASSIFICATION RESULTS ON IEMOCAP DATABASE FOR LEAVE-ONE-SPEAKER-OUT-CROSS VALIDATION

| System | UAR | WAR |
|---|---|---|
| MFCC-HMM [16] | 50.69% | - |
| Hierarchical binary decision tree [4] | 58.46% | 56.38% |
| Eigen channel factor-SVM [5] | 55.84% | 55.88% |
| MFCC-GMM (baseline) [5] | 54.11% | - |
| Speaker Normalization [21] | 56.73% | - |
| Lexical Normalization [21] | 55.37% | - |
| Baseline 1:emo-ISO9-SVM | 56.75% | 56.28% |
| Baseline 2:MFCC-Supervector-SVM *(512 GMM)* | 55.50% | 53.93% |
| Baseline 3:MFCC-i-Vector-SVM (*256 GMM i-vector length-75)* | 56.95% | 56.20% |
| **Proposed i-Vector-GPLDA** *(256 GMM i-vector length 92)* | **58.20%** | **55.92%** |

It can be seen from Table I that the proposed i-vector GPLDA system achieved an unweighted average recall (UAR) of 58.2% which outperformed all three baseline systems and matched the best reported results in the literature on this database. Specifically, comparing the results from baseline 2, baseline 3 and the proposed system, it can be seen that the introduction of the i-vectors leads to a performance improvement (1.45% absolute) and replacing the SVM back-end with a GPLDA back-end led to a further improvement (1.25% absolute).

The best performing system evaluated with only acoustic features on this database was reported in [4] and its performance is matched by the performance of the proposed system.

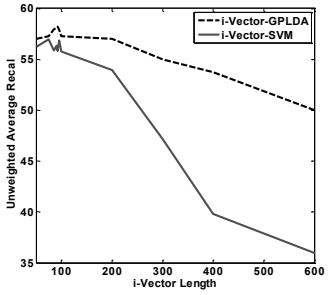


Fig. 2 Variation of UAR with varying i-vector length.

Both the hierarchical tree based system reported in [4] as well as the proposed system benefit from tuning the system parameters. However, the proposed system does not require an additional phase (of preliminary experiments) to determine the structure of the hierarchical tree [4]. In addition to improved classification performance, it was also observed that the GPLDA back-end was more robust to i-vector dimension tuning. Both the GPLDA based and SVM based systems showed an optimum performance around the region of i-vector length from 75 to 95 while the UAR reduced for higher

and lower i-vector lengths gradually from the optimum region. This behavior is most likely due to the fact that the use of higher dimensional i-vectors may lead to the inclusion of additional information unrelated to emotions, while the use of low dimensional i-vectors may lead to loss of information required to distinguish between emotional classes, both of which lead to lowered performance.

However, in comparison to the SVM back-end, the GPLDA back-end was significantly less sensitive to i-vector dimensionality and in particular having a somewhat higher dimensional i-vector had negligible impact on the system (Fig. 2).

This robustness to i-vector dimensionality may be because the GPLDA back-end weights the more discriminative i-vector dimensions higher than the less discriminative dimensions. Thus, even when the i-vectors are noisy, the GPLDA perform an implicit feature selection through the weighting process leading to better performance compared to SVMs.

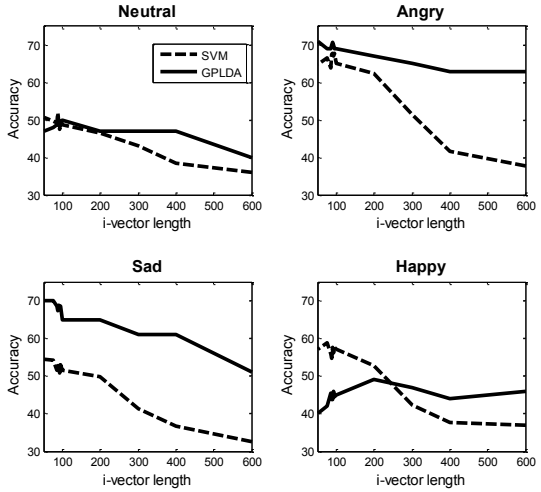


Fig. 3 Effect of i-vector length on average classification accuracies of different emotional categories

Fig. 3 compares the performance of the proposed GPLDA system on the 4 individual emotion category recall rates to that of the SVM based back-end (Baseline-3) as a function of i-vector dimensionality.

## V. CONCLUSIONS

The proposed i-vector GPLDA system offers state of-the-art speech emotion classification performance on the IEMOCAP database. Specifically, our results show that a GPLDA back-end can outperform a SVM based back-end in the context of emotion classification. Finally, our experiments demonstrate that a GPLDA backend is significantly less sensitive to i-vector dimensionality when compared to SVM back-ends. This is of particular significance since the i-vector dimensionality is a system hyperparameter that is normally tuned on some development data and consequently the optimal value is likely to vary between databases. However, if the back-end is not sensitive to the i-vector dimensionality, hyper-parameter tuning to determine the optimal dimensionality may be avoided or minimized.

## REFERENCES

[1] B. Schuller and A. Batliner, *Computational paralinguistics: emotion, affect and personality in speech and language processing*. John Wiley & Sons, 2013.

[2] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition*, vol. 44, no. 3, pp. 572–587, 2011.

[3] V. Sethu, J. Epps, and E. Ambikairajah, "Speech based emotion recognition," in *Speech and Audio Processing for Coding, Enhancement and Recognition*. Springer, 2015, pp. 197–228.

[4] C.-C. Lee, E. Mower, C. Busso, S. Lee, and S. Narayanan, "Emotion recognition using a hierarchical binary decision tree approach," *Speech Communication*, vol. 53, no. 9, pp. 1162–1171, 2011.

[5] M. Li, A. Metallinou, D. Bone, and S. Narayanan, "Speaker states recognition using latent factor analysis based eigenchannel factor vector modeling," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. IEEE, 2012, pp. 1937–1940.

[6] S. Chen, Q. Jin, X. Li, G. Yang, and J. Xu, "Speech emotion classification using acoustic features," in *Chinese Spoken Language Processing (ISCSLP), 2014 9th International Symposium on.* IEEE, 2014, pp. 579–583.

[7] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 4, pp. 788–798, 2011.

[8] C. S. Greenberg, D. Bansé, G. R. Doddington, D. Garcia-Romero, J. J. Godfrey, T. Kinnunen, A. F. Martin, A. McCree, M. Przybocki, and D. A. Reynolds, "The NIST 2014 speaker recognition i-vector machine learning challenge," in *Odyssey: The Speaker and Language Recognition Workshop*, 2014.

[9] R. Xia and Y. Liu, "Using i-vector space model for emotion recognition," in *Thirteenth Annual Conference of the International Speech Communication Association,* 2012.

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

[11] D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of i-vector Length Normalization in Speaker Recognition Systems." in *Interspeech*, 2011, pp. 249–252.

[12] S. Ioffe, "Probabilistic linear discriminant analysis," in *Computer Vision–ECCV 2006.* Springer, 2006, pp. 531–542.

[13] T. Hastie and R. Tibshirani, "Discriminant analysis by Gaussian mixtures," *Journal of the Royal Statistical Society. Series B (Methodological),* pp. 155–176, 1996.

[14] 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.* IEEE, 2007, pp. 1–8.

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

[16] A. Metallinou, S. Lee, and S. Narayanan, "Decision level combination of multiple modalities for recognition and analysis of emotional expression," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on.* IEEE, 2010, pp. 2462–2465.

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

[18] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *Signal Processing Letters, IEEE*, vol. 6, no. 1, pp. 1–3, 1999.

[19] S. O. Sadjadi, M. Slaney, and L. Heck, "MSR identity toolbox v1. 0: A MATLAB toolbox for speaker recognition research," *Speech and Language Processing Technical Committee Newsletter,* 2013.

[20] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: an update," *ACM SIGKDD explorations newsletter,* vol. 11, no. 1, pp. 10–18, 2009.

[21] S. Mariooryad and C. Busso, "Compensating for speaker or lexical variabilities in speech for emotion recognition," *Speech Communication*, vol. 57, pp. 1–12, 2014.