# iVectors for Continuous Emotion Recognition

Paula Lopez-Otero, Laura Docio-Fernandez, and Carmen Garcia-Mateo

Multimedia Technologies Group (GTM), AtlantTIC Research Center
E.E. Telecomunicación, Campus Universitario de Vigo S/N
36310, Vigo, Spain
{plopez,ldocio,carmen,cardenal}@gts.uvigo.es

**Abstract.** This work proposes the use of the iVectors paradigm for performing continuous emotion recognition. To do so, a segmentation of the audio stream with a fixed-length sliding window is performed, in order to obtain a temporal context which is enough for capturing the emotional information of the speech. These segments are projected into the iVectors space, and the continuous emotion labels are learnt by canonical correlation analysis. A voice activity detection strategy is incorporated to the system in order to ignore the non-speech segments, which do not provide any information about the emotional state of the speaker, and to recreate a real-world scenario. Results on the framework of the Audiovisual Emotion Challenge (AVEC) 2013 show the potential of this approach for the emotion recognition task, obtaining promising results as well as using low-dimensional data representation.

**Keywords:** iVectors, continuous emotion recognition, arousal, valence

## 1 Introduction

Emotion recognition is a task consisting on extracting information about the speaker's emotional state. The emotion recognition field is gaining interest for different real-world applications. The emotional aspects of human-computer interaction started to receive some attention in the last years, becoming a huge concern [15]; this interest is caused by the importance of expressivity when dealing with a computer interface, as the interface should be able to detect the emotional state of the user and adapt its behaviour according to it [4]. The use of emotion recognition for monitoring users' reaction to advertisement is nowadays a reality, and it also has paramount importance in the entertainment industry, either for the development of toys or videogames [14] or for the development of serious games for aiding people with problems to understand social signs [1].

Emotion recognition can be carried out in two different ways: one is the discrete emotion recognition task, that consists on detecting a set of given emotions, and the other one is the continuous emotion recognition task, in which the continuous values of the affect variables have to be estimated [9]. The two problems are closely related, as a discrete emotion has a correspondence with concrete values of emotional dimensions; in [16], it is said that the different dimensions of affect lie in different angles of a circle, and the angle inside this circle depends on

two different emotional dimensions, namely arousal, which measures the degree of excitation, and valence, which measures the degree of pleasantness.

Emotion recognition systems can be usually split in three stages, namely feature extraction, modelling and prediction of the emotional level. On the first stage, the features used in emotion recognition can be divided in two big groups according to their time span: low level descriptors (LLD) or instantaneous descriptors are extracted for each time frame, such as Mel-frequency cepstral coefficients, loudness, zero crossing rate, jitter or shimmer; and functionals or global descriptors are computed using the LLD extracted for the whole audio signal or for an audio segment covering several audio frames, such as the mean, standard deviation, quartile, flatness or skewness, among others.

The modelling stage of an emotion recognition system must obtain a representation of the speech that reflects the emotional information. Depending on the features used, different modelling approaches can be found in the literature. When using functionals, it is common to model the speech using those features directly or applying feature selection strategies [8]. When dealing with LLD, different techniques can be borrowed from other speech recognition tasks, such as supervised and unsupervised subspace learning techniques. The use of such techniques in discrete emotion recognition is straightforward, but their application to continuous emotion recognition has some issues. First, there is not a discrete number of emotions, so the training stage of a supervised learning strategy cannot be carried out directly; this issue can be partially solved by quantizing the emotional dimensions [17]. Moreover, an instantaneous value of the emotion must be estimated, but it is not possible to apply such learning techniques at a frame rate, as the modelling of speech through these techniques usually requires a context longer than a few milliseconds. A windowing of the speech can be done before the modelling stage in order to obtain a longer context [17].

The prediction of the emotional level is usually carried out using machine learning techniques such as support vector machines (SVM) or random forests in the discrete case, while in the continuous case strategies such as support vector regression (SVR) or canonical correlation analysis [10] can be used.

In this work, we present a continuous emotion recognition system as a continuation of our previous research on speech modelling in continuous emotion recognition. In [17], we successfully applied the eigen-voices approach to emotion recognition, and in this work we propose the use of the iVector modelling technique. The iVector paradigm is considered state-of-art in different tasks such as speaker recognition and verification [5] or language identification [6] due to its potential for representing audio in a speaker and channel independent way in a low dimensional subspace, which are desirable qualities in emotion recognition. The iVector representation was used in [19] for discrete emotion recognition, but in this work we present a system that permits the use of this representation for the continuous task. The validity of this approach is assessed in the framework of the AVEC 2013 affect recognition sub-challenge [13], which consisted on estimating the continuous levels of arousal and valence in a set of recordings featuring different speakers and channel conditions.

The rest of this paper is organized as follows: Section 2 presents the proposed system for performing continuous emotion recognition; Section 3 describes the experimental framework used to assess the validity of the proposed system; the experimental settings are described in Section 4; experimental results are discussed in Section 5; and Section 6 summarizes the conclusions and future work.

## 2 Proposed continuous emotion recognition system

Figure 1 presents an overview of the proposed approach for performing continuous emotion recognition, whose different blocks are described in detail in the rest of this Section. The emotional levels on $E$ different dimensions have to be estimated by this approach, which, as mentioned in the introduction, can be divided in three stages: feature extraction, accompanied by a segmentation step, iVector modelling and estimation of the emotional level. This system is similar to the one presented in [17], as the feature extraction and segmentation procedures, as well as the approach for predicting the emotional level, are almost the same; these two systems differ in the modelling of the speech segments, which in this case is carried out using the iVector paradigm.

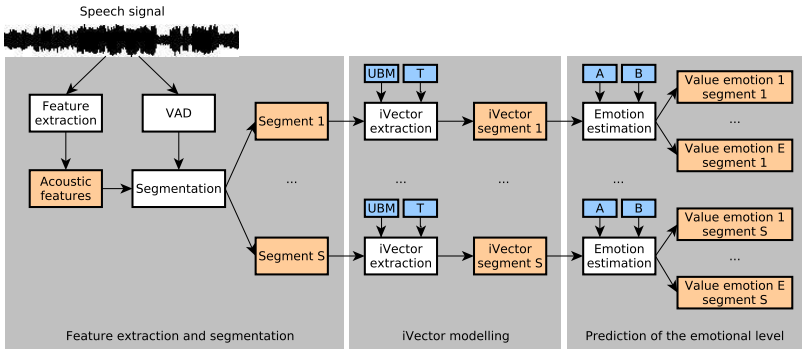


**Fig. 1.** Continuous emotion recognition system

### 2.1 Feature extraction and segmentation

The first step consists on extracting vectors of $F$ features from the audio signal. Acoustic features represent a temporal context which usually ranges from 20 to 100 ms; it is not possible to identify an emotion with such a short time span, a bigger context is necessary. Thus, to obtain vectors that represent a bigger context, the audio is segmented using a fixed-length sliding window, obtaining

a set of $S$ segments. In this way, the segments of audio can be represented by means of Gaussian mixture models (GMMs).

It must be noted that the audio signals may contain non-speech parts, which act as a nuisance in the continuous emotion recognition system, i.e. the non-speech parts contribute in a negative way to this procedure, as they do not hold any information about the speaker's emotional state. Hence, it is important to perform voice activity detection (VAD) on the audio signals. To do so, the approach described in [2] was used, which uses a linked-HMM architecture and robust features that are independent of the signal energy, making this strategy robust to noise. An audio segment is considered to contain speech if at least the 50% of its duration was labelled as speech by the VAD strategy.

### 2.2 iVector modelling

The procedure performed at the previous stage of this system allows the use of iVectors for modelling the acoustic information in a low-dimensionality space. Given a Universal Background Model (UBM) with $N$ mixtures, this UBM is adapted to the segments extracted from the training files using Maximum a Posteriori (MAP) adaptation, and the means of the resulting Gaussian Mixture Model (GMM) are concatenated in order to obtain a Gaussian mean supervector for each segment. As we want to avoid the effects of speaker and channel variability, the iVector technique is applied to the Gaussian mean supervectors. This technique defines a low-dimensional space, named total variability space, in which the speech segments are represented by a vector of total factors, namely iVector [5]. A Gaussian mean supervector $\mathbf{M}$ is decomposed as follows:

$$\mathbf{M} = \mathbf{m} + \mathbf{T}\mathbf{w} \tag{1}$$

where $\mathbf{m}$ is the speaker and channel independent supervector, $\mathbf{T}$ is a low-rank total variability matrix, and $\mathbf{w}$ is the iVector corresponding to the Gaussian mean supervector. In this training stage, the matrix $\mathbf{T}$ is trained as described in [11].

Once the total variability matrix $\mathbf{T}$ is obtained, this matrix and the UBM can be used to extract iVectors from the acoustic features corresponding to the different speech segments.

### 2.3 Prediction of the emotional level

The iVector modelling strategy described above results on an iVector for each speech segment, which represents all the emotional dimensions at the same time. Estimated values of the emotional levels of the different dimensions must be extracted from these vectors; to do so, canonical correlation analysis is applied. This technique finds basis vectors for two sets of variables (on the one hand, the iVectors, and on the other hand, the groundtruth emotional levels) such that the correlations between the projection of the variables onto these basis vectors are mutually maximized [10]. After training the canonical correlation projection,

two matrices are obtained which are used to project the iVectors, obtaining as a result the estimated emotional levels.

Two different strategies can be followed at this point: a projection can be learnt for each emotional dimension, or a single projection for all the emotional dimensions can be obtained. The first approach does not take into account the correlation between the different emotional dimensions, but the second approach takes advantage of this correlation, which might be due to the emotions themselves or might be caused by the rater that labelled the different emotional dimensions.

It must be noted that, as commented in Section 2.1, the non-speech segments are not used either for training or testing. Thus, there are parts of the audio files whose emotional level is not estimated, as they do not have a corresponding iVector, but it might be necessary to assign them a value for evaluation purposes. To solve this situation, the mean value of the emotional dimension, computed over the training labels, is assigned to the non-speech segments.

## 3 Experimental Framework

The framework of the AVEC 2013 affect recognition sub-challenge (ASC) was used to evaluate the proposed technique for continuous emotion recognition. This task consists on the continuous recognition of the emotional dimensions valence and arousal, and these dimensions have to be predicted for every time instant of the recording. Both arousal and valence values range from -1 to 1.

The data used in these experiments is a subset of the audio-visual depressive language corpus (AVDLC) [13]. The speakers were recorded in diverse quiet locations using a laptop and a headset. The database is organized in three partitions of 50 recordings each, which are summarized in Table 1. Each recording features one speaker (either male or female), and there can be several recordings per speaker, with a time separation of two weeks between different recordings. The subjects' age ranged between 18 and 63 years (mean 31.5).

**Table 1.** Summary of the datasets used in the experiments.

| Set | Total duration | Min duration | Max duration |
|---|---|---|---|
| Training | 13 h 17 min | 8 min 5 s | 27 min 20 s |
| Development | 13 h 5 min | 14 min 20 s | 23 min 55 s |
| Testing | 12 h 59 min | 5 min 15 s | 23 min 57 s |

The recordings were power-point guided, indicating the speaker what to do at each moment. These tasks consisted in reading excerpts of novels and fables, singing, telling stories from the speaker's past, making up a story applying the Thematic Aperception Test (TAT), and sustained vowel phonation. These tasks try to provoke different feelings on the speakers such as happiness, by talking

about their best present ever, or sadness, by talking about their saddest memory of their childhood.

The recordings included in the database were manually labelled by a team of 23 raters. Each recording was annotated by a single rater, which used a joystick to instantaneously register the level of arousal or valence (the two dimensions were annotated separately, not at the same time). In order to address the intra-annotator variability, all the raters were asked to annotate a reference video, and these annotations were used to create models that compensated that variability. The annotations were binned in temporal units of time of the same duration, which in this case was 1/30 seconds (i.e. equal to the video frame rate).

Figure 2 shows the distribution of arousal and valence in the training and development datasets (the testing dataset is not included due to the unavailability of its corresponding groundtruth labels). As shown in this Figure, the most probable value of arousal and valence is 0, which means neutral arousal or valence. This is due to the fact that, during the recordings, there are long silence periods which the raters labelled as neutral. It can also be observed in this Figure that values close to -1 and 1 are not very likely to appear.

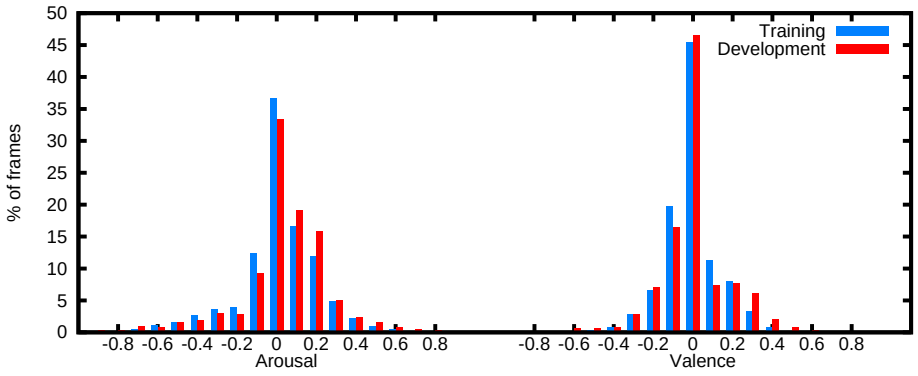


**Fig. 2.** Distribution of arousal and valence values in the training and development datasets

The evaluation metric used to measure the performance of continuous emotion recognition approaches in this framework is the Pearson's correlation coefficient between the groundtruth labels and the estimated labels, averaged over all sessions and dimensions. It must be noted that the absolute value of the correlation coefficient is used, as a negative correlation is a correlation anyway. This fact was not mentioned in the description of the AVEC 2013 ASC challenge [13], but was mentioned afterwards when the evaluation scripts were released.

## 4 Experimental settings

In this continuous emotion recognition system, 16 MFCCs plus delta and derivatives are used, as these features are a common choice for emotional speech analysis [18], leading to feature vectors of dimension 48. The use of MFCCs is also supported by previous research in [17], where the best emotion recognition results were obtained using MFCCs. These features were extracted every 10 ms using a 40 ms Hamming window and mean and variance normalization is applied. The window used to perform the audio segmentation has a length of 3 s with 1 s of overlap. The number of mixtures of the GMM was set to 256 after observing that the influence of this value in the experimental results was negligible. Feature extraction was done using the OpenSMILE feature extraction software [7], and all the tasks that involved iVector training and extraction were performed using the ALIZE 3.0 toolkit [12].

The training partition of the AVEC 2013 ASC database was used to perform the training of the system (UBM, total variability matrix and canonical correlation projection) as well as to adjust the free parameters of the system, i.e. the dimension of the iVectors. The development partition was used to test the performance of the proposed system, and the testing data was discarded due to the unavailability of their corresponding groundtruth labels.

The validity of the iVector approach for emotion recognition was assessed by comparing the proposed system with the same system without applying the iVectors paradigm, i.e. a GMM supervector (GSV) approach [3]. In this GSV approach, the audio segments were represented by mean supervectors $\mathbf{m}$ of dimension $N \cdot F = 256 \cdot 48 = 12288$, according to the notation in Section 2.2.

It must be noted that the rate of the groundtruth labels is higher than the rate of the iVectors (there is one iVector every two seconds while there is a groundtruth label every 1/30 seconds); in order to make them have the same rate, the mean value of the emotional dimension on the time span of the iVector is used as the groundtruth label to train the canonical correlation basis vectors. In the test data, as an emotional level is obtained every 2 seconds, the obtained level is replicated so it fits the rate of the groundtruth labels (i.e. every label is repeated $\frac{2}{1/30} = 60$ times).

## 5 Experimental results

The starting point of the experimental validation of the proposed strategy consisted on performing continuous emotion recognition using different dimensions of the iVectors. To do so, the Pearson's correlation coefficient between the groundtruth training labels and the estimated ones was computed for different iVector dimensions, as presented in Table 2. This Table shows that the highest correlation was obtained when using iVectors of dimension 25. It must be noted that these results were obtained when using the manual VAD of the audio signals, in order to avoid the nuisance generated by errors on the VAD stage.

A straightforward experiment to prove the validity of the iVector technique consists on comparing the results when applying this approach and when not

**Table 2.** Pearson's correlation coefficient obtained on the training dataset with different iVector dimensions using manual VAD.

| Dimension | Arousal | Valence | Average |
|---|---|---|---|
| 25 | **0.1963** | 0.1862 | **0.1912** |
| 50 | 0.1859 | **0.1905** | 0.1882 |
| 100 | 0.1847 | 0.1892 | 0.1870 |
| 150 | 0.1846 | 0.1868 | 0.1857 |
| 200 | 0.1863 | 0.1852 | 0.1857 |

applying it. Thus, results obtained with the GSV approach were compared with those obtained when using the iVector approach. Table 3 supports the validity of the iVector modelling, as the Pearson's correlation coefficient of the estimation of arousal and valence obtained with the iVectors representation is higher than that obtained when modelling the speech segments with the GSV approach. Another issue that must be noticed is the dramatic reduction of the dimensionality of the vectors: while the dimension of the iVectors was 25, the GSV approach used vectors of dimension 12288.

**Table 3.** Pearson's correlation coefficient obtained on the development dataset with and without using the iVector modelling.

| VAD | Approach | Dimension | Arousal | Valence | Average |
|---|---|---|---|---|---|
| Manual | iVectors | 25 | **0.2041** | **0.1697** | **0.1869** |
| | GSV | 12288 | 0.1717 | 0.1299 | 0.1508 |
| Automatic | iVectors | 25 | **0.1846** | **0.1575** | **0.1711** |
| | GSV | 12288 | 0.1621 | 0.1264 | 0.1443 |
| - | Eigen-emotions | 50 | 0.1721 | 0.1404 | 0.1562 |

The last step of this experimental validation of the iVector modelling for emotion recognition was to apply an automatic VAD strategy to the development segments, in order to observe the impact of making errors in the detection of speech. The VAD strategy described in Section 2.1 was used for that purpose; comparing the manual VAD to the automatic VAD results obtained with such strategy, a missed speaker time of 18.6% and a false alarm speaker time of 1.8%, with respect to the scored speaker time, were achieved. Table 3 shows that the system presents some sensitivity to the VAD errors, as the Pearson's correlation coefficient is reduced by 0.015 when using iVectors and by 0.007 when using the GSV approach. These errors are due to the missed speaker time, as there are speech segments whose emotional dimensions are not being estimated because they were labelled as non-speech.

Table 3 also shows the Pearson's correlation coefficient achieved with another subspace projection-based approach available in the literature, which used the same experimental framework. This approach, namely eigen-emotions [17], shows

similar results to those obtained with the GSV technique, but it is outperformed by the iVectors representation.

## 6 Conclusions and future work

This work proposed the use of iVectors on the continuous emotion recognition task, due to the ability of this paradigm to get rid of the speaker and channel variabilities. The experimental results obtained in the framework of the AVEC 2013 affect recognition sub-challenge showed an improvement on the emotion recognition results when using the iVector paradigm, as well as a huge dimensionality reduction of the feature vectors used to represent the speech segments. Hence, the experimental results suggest that the success achieved by the iVector representation in different speech technologies tasks is extensible to the emotion recognition field as well. Nevertheless, due to the temporary unavailability of the groundtruth labels of the test data used in the AVEC 2013 evaluation, it was not possible to compare these results to those obtained on the test data using other systems, but a more extensive analysis of this approach will be performed whenever these labels are available.

An automatic strategy to discriminate speech and non-speech was applied in order to assess the performance of the proposed emotion recognition approach in a realistic scenario; this procedure resulted in a slight reduction of the Pearson's correlation coefficient due to the errors on the voice activity detection module, specially to the missed speech errors. Better strategies for voice activity detection must be developed and incorporated to this emotion recognition system in order to overcome this reduction of performance.

## Acknowledgements

This work has been supported by the European Regional Development Fund, the Galician Regional Government (CN2011/019, 'Consolidation of Research Units: AtlantTIC Project' CN2012/160), the Spanish Government (FPI grant BES-2010-033358 and 'SpeechTech4All Project' TEC2012-38939-C03-01) and 'TecAnDAli Research Network' of the Consellería de Educación e Ordenación Universitaria, Xunta de Galicia.

## References

1. Barakova, E.I., Lourens, T.: Expressing and interpreting emotional movements in social games with robots. Personal Ubiquitous Computing 14(5), 457–467 (2010)
2. Basu, S.: A linked-HMM model for robust voicing and speech detection (2003)
3. Chen, Y., Xie, J.: Emotional speech recognition based on SVM with GMM supervector. Journal of Electronics (China) 29(3), 339–344 (2012)
4. Cowie, R., Douglas-Cowie, E., Karpouzis, K., Caridakis, G., Wallace, M., Kollias, S.: Recognition of emotional states in natural human-computer interaction. Multimodal User Interfaces pp. 119–153 (2008)

5. Dehak, N., Kenny, P.J., Dehak, R., Dumouchel, P., Ouellet, P.: Front end factor analysis for speaker verification. IEEE Transactions on Audio, Speech and Language Processing (2010)
6. Dehak, N., Torres-Carrasquillo, P., Reynolds, D., Dehak, R.: Language recognition via i-vectors and dimensionality reduction. In: Proceedings of Interspeech. pp. 857–860 (2011)
7. Eyben, F., Wöllmer, M., Schuller, B.: OpenSMILE - the Munich versatile and fast open-source audio feature extractor. In: Proceedings of ACM Multimedia (MM). pp. 1459–1462 (2010)
8. Gosztolya, G., Busa-Fekete, R., Tth, L.: Detecting autism, emotions and social signals using AdaBoost. In: INTERSPEECH. pp. 220–224 (2013)
9. Gunes, H., Pantic, M.: Automatic, dimensional and continuous emotion recognition. International Journal of Synthetic Emotions 1(1), 68–99 (2010)
10. Hardoon, D.R., Szedmak, S., Szedmak, O., Shawe-taylor, J.: Canonical correlation analysis; an overview with application to learning methods. Tech. rep. (2007)
11. Kenny, P., Boulianne, G., Dumouchel, P.: Eigenvoice modeling with sparse training data. IEEE Transactions on Speech and Audio Processing 13(3), 345–354 (2005)
12. Larcher, A., Bonastre, J., Fauve, B., Lee, K., Levy, C., Li, H., Mason, J., Parfait, J.: ALIZE 3.0 - open source toolkit for state-of-the-art speaker recognition. In: Proceedings of Interspeech. pp. 2768–2772 (2013)
13. M.Valstar, Schuller, B., Smith, K., Eyben, F., Jiang, B., Bilakhia, S., Schnieder, S., Cowie, R., Pantic, M.: AVEC 2013 - the continuous audio/visual emotion and depression recognition challange. In: Proceedings of the 3rd International Audio/Visual Emotion Challenge and Workshop (AVEC'13) (2013)
14. Oudeyer, P.: The production and recognition of emotions in speech: features and algorithms. International Jorunal of Human-Computer Studies 59, 157–183 (2003)
15. Picard, R.W.: Toward computers that recognize and respond to user emotion. IBM Syst. J. 39(3-4), 705–719 (Jul 2000)
16. Russell, J.: A circumplex model of affect. Journal of Personality and Social Psychology 39, 1161–1178 (1980)
17. Sánchez-Lozano, E., Lopez-Otero, P., Docio-Fernandez, L., Argones-Rúa, E., Alba-Castro, J.L.: Audiovisual three-level fusion for continuous estimation of Russell's emotion circumplex. In: Proceedings of the 3rd ACM International Workshop on Audio/Visual Emotion Challenge. pp. 31–40. AVEC '13, ACM (2013)
18. Sato, N., Obuchi, Y.: Emotion recognition using Mel-frequency cepstral coefficients. Information and Media Technologies 2(3), 835–848 (2007)
19. Xia, R., Liu, Y.: Using i-vector space model for emotion recognition. In: INTERSPEECH. ISCA (2012)