Continuous Emotion Tracking using Total Variability Space

Hossein Khaki, Engin Erzin

Department of Electrical and Electronics Engineering, Koç University, Istanbul, Turkey

Hkhaki13,eerzin@ku.edu.tr

Abstract

Automatic continuous emotion tracking (CET) has received increased attention with expected applications in medical, robotic, and human-machine interaction areas. The speech signal carries useful clues to estimate the affective state of the speaker. In this paper, we present Total Variability Space (TVS) for CET from speech data. TVS is a widely used framework in speaker and language recognition applications. In this study, we applied TVS as an unsupervised emotional feature extraction framework. Assuming a low temporal variation in the affective space, we discretize the continuous affective state and extract i-vectors. Experimental evaluations are performed on the CreativeIT dataset and fusion results with pool of statistical functions over mel frequency cepstral coefficients (MFCCs) show a 2% improvement for the emotion tracking from speech.

Index Terms: total variability space, i-vector, continuous emotion tracking, Gaussian mixture regression.

10.21437/Interspeech.2015-324

1. Introduction

Automatic continuous emotion tracking (CET) aims to automatically estimate the level of the emotion or affective state from the speech, video, and physical signals of a person. This goal recently has received increasing attention, and has the potential to define applications in medical, robotic, education, and commercial areas such as affective human machine interaction, public speaking training, autism monitoring, and voice response systems [1, 2]. In psychology literature, various affective space representations have been employed to model emotion [3]. In a high level categorization, affective space representations are divided into discrete and continuous models.

Discrete models often have separate categories or combinations of several basic emotions, such as: neutral, happiness, sadness, surprise, fear, anger, and disgust. However, in continuous models, a vector in a multi-dimensional space, affective space, defines the emotion along the time index. One of the widely used affective spaces is the Activation-Valence-Dominance (AVD) space, which describes the intensity, level of pleasure, and amount of control of the emotion, respectively. These are referred to as attributes or dimensions of the affective space [2, 4].

CET is a regression from feature space to affective space along the time index. The features can be extracted from speech, video or other physical signals of the person. In this paper, we look at the speech channel, but the methods can be combined with other channels in a multimodal framework. Along time, CET can be applied in a frame level or a window level resolution [4]. In the frame level, the durations of the units are roughly less than 0.5 sec; hence, a good time precision can be obtained and the features can be defined over frames. However, from psychological point of view [5], we know that the overall duration of the emotions is 0.5 to 4 sec. Hence, the window level methods of CET where a single vector represents the affective state over a temporal window of several seconds are more suitable.

Obtaining high level feature sets over a number of frames in a window of time is one of the challenges that we address in this paper. This challenge is also addressed as sequence summarization in action recognition by video processing [6]. In [4], for the window level resolution, they used a variety of statistical functionals followed by principal component analysis (PCA) for dimensionality reduction and trained a dynamic Gaussian mixture regression (GMR) to track the AVD attributes. In this paper, we use Total Variability Space (TVS) to summarize the frames of a temporal window.

TVS is among the state of the art systems for speaker verification. It defines two spaces, namely speaker variability and session variability, and extracts the features that are more informative for speaker verification. Recently it has received increased attention in other applications such as age estimation [7] and discrete emotion classification [8]. The reported results in [8] show that TVS can model the emotion variability as well as speaker variability. Although using TVS for discrete emotion classification is just a generalization of TVS for speaker verification, it cannot be directly used for CET. TVS based feature extraction should be defined over temporal windows where in each window sufficient statistical information is needed to define TVS. In this paper, we present a new framework for CET by defining TVS-based feature extraction over temporal windows and by mapping the i-vectors of the speech signal to the continuous affective space dimensions.

The rest of the paper is organized as follows: in section 2, we define the proposed CET system. In section 3, we introduce the datasets in use, system setup, and experimental results. Conclusions are given in section 4.

2. Continuous Emotion Tracking

We use frame level and window level resolution for the baseline systems. For the frame level system, as described in [4], a 16.66 msec frame of speech data with 50% overlap is used to extract MFCC features and then GMR is applied to estimate the affective state. For the window level system, we summarized the features over overlapping temporal windows using a summarization function, $\mathcal{F}: \mathcal{R}^{m*N} \rightarrow \mathcal{R}^{k}$, in which m is the dimension of features for each frame, N is the number of frames over a time window, and k is the dimension of the summarized features. In the literature, a variety of statistical functions such as mean, standard deviation, median, minimum, maximum, range, skewness, kurtosis, the lower and upper

quantiles (corresponding to the 25th and 75th percentiles) and the interquantile range followed by PCA to reduce the dimension, is used for summarization purpose [4]. Another option is to concatenate all the frame level features of the window and then apply PCA to reduce the dimension. Similar to the frame level system, GMR maps the window level features to the affective state. Unlike the frame level resolution the estimated affective states in the window level system do not have the same sampling rate as the original annotation data. So we interpolate the affective dimensions at the frame rate and calculate the mean of the affective dimensions over the overlapping windows.

In our CET system, speech signal feature summarization is performed using TVS. In section 2.2, we first describe TVS in terms of emotion and state variability, then we present the discretization method in section 2.3, and finally we present the whole model for CET in section 2.3.

2.1. Total Variability Space

TVS is a widely used framework in speaker and language recognition applications, and refers to speaker, language and session variability spaces. In the TVS first a Gaussian mixture model models the distribution of the data:

$$P(\mathcal{D}) = \sum_{i=1}^{M} \omega_i \mathcal{N}\left(\mathcal{D}; \underline{\mu_i}, \boldsymbol{\Sigma_i}\right), \quad (1)$$

where $\mathcal{D}$ is the speech feature space, ω_i, $\underline{\mu_i}$, and Σ_i are the weight, mean vector, and covariance matrix of the i'th Gaussian mixture respectively and M is the total number of mixtures. Then the super-vector, which is the concatenation of mean vectors $\underline{\mu_i}$, is mapped to a lower dimensional TVS space as, $\underline{\mu} = \underline{m} + T\underline{w}$, where $\underline{\mu}$ is the super-vector, $\underline{m}$ is a representor, usually the concatenated mean vectors of the universal background model (UBM), T matrix represents the TVS basis, and $\underline{w}$ is the extracted feature vector, which is known as i-vector in verification literature. The details of calculating of T matrix are given in [9].

In the emotion tracking problem, we can similarly define emotion variability and state variability spaces. The primary space of the continuous emotion is the AVD dimensions of the affective space. The secondary space stands for any other differences between the recordings of the same affective dimensions (such as the affective variability over the speakers, environment or context). By combining these two spaces we can define the TVS for the emotion tracking problem.

Recently i-vector has been used for feature extraction in discrete emotion classification [8]. In [8] they used the whole utterance of a single emotion, as $\mathcal{D}$ in (1), to generate $\underline{\mu}$ and $\underline{w}$, then performed classification task. However in our framework, we target the CET problem in the AVD space. Hence, we need a population of data representing an affective state value to extract the i-vector representation in (1). Therefore, we used overlapping temporal windows for speech signal and performed a window level discretization of frame level affective dimensions. In the next subsection we describe the discretization of AVD space.

2.2. Discretization of Continuous Emotion

Continuous emotion is represented in AVD dimensions. These three dimensions build a continuous space where affective state takes values from this space at each time frame. These continuous attributes help to track the emotion of a person along time. Assuming a slowly changing affective state, we modeled AVD dimensions as a single point to represent the emotional content of a temporal window. Then overlapping temporal windows were utilized for the discretization of the continuous emotion space. The AVD point that represents a temporal window can be the average, median or any other statistics of the AVD dimensions in the corresponding window. A distortion metric for the discretization of continuous emotion space is defined as:

$$e_d = \sqrt{\frac{1}{NL_w} \sum_{i=0}^{N-1} \sum_{t=i*L_f}^{i*L_f+L_w-1} (a_t - \alpha_i)^2}, \quad (2)$$

where a_t is one of the AVD dimensions at time t, α_i is the discrete representation of a_t in the i-th window with L_w samples, L_f is the time shift of the window in samples, and N is the total number of windows. We use α_i as the mean attribute of the window i, so the distortion e_d can be interpreted as the standard deviation of attributes. We can choose the window length L_w based on an accepted tolerance of discretization distortion. After the discretization, we are able to extract the $\underline{\mu}$ and $\underline{w}$ vectors for each window.

2.3. Continuous Emotion Tracking System

Our CET system is based on the TVS and GMR. Figure 1 depicts a block diagram of the proposed system model. In Figure 1(a) we train the UBM and extract the T matrix over training data that contains emotion and state variability. Then in Figure 1(b), we calculate the i-vector, $\underline{w_i}$, for the i'th window of the speech data. In Figure 1(c) discretization of the affective dimensions is performed over overlapping windows based on (2). Discrete affective dimension values are taken as emotion labels of the windows. Then we train the GMR model with the i-vectors and corresponding labels. In Figure 1(d) we use the GMR model and estimate the emotion labels for a test i-vector, which is extracted from the test speech signal. After the GMR, the estimated attributes, $\widehat{\alpha}_i$, are low-pass filtered for smoothing as in [9]. To achieve the same time resolution we interpolate the smoothed values as stated in section 2.1.

Although the mean square error between the estimated and mean of the annotators' AVD dimensions is a possible evaluation metric, we choose to evaluate the performance with the correlation metric since variation of the AVD dimensions is more important than the exact values.

3. Experimental Results

3.1. Datasets

To train the UBM and extract the T matrix we use four different datasets, which contains emotion and state variability. Since training the UBM is unsupervised, we do not need annotation information. Hence, we use IEMOCAP [10], Vera am Mittag (VAM) [11], IS10 Paralinguistic Challenge [12], and Interspeech 2009 Emotion Challenge (IS09) [13] datasets for the training of the UBM. These datasets contain more than 30 hours of affective speech.

On the other hand, we need a dataset, which contains continuous annotation for testing the i-vector representation and CET. We use the CreativeIT dataset [14], which includes improvised dyadic interactions with continuous annotation of AVD dimensions. This dataset contains 8 sessions, where in

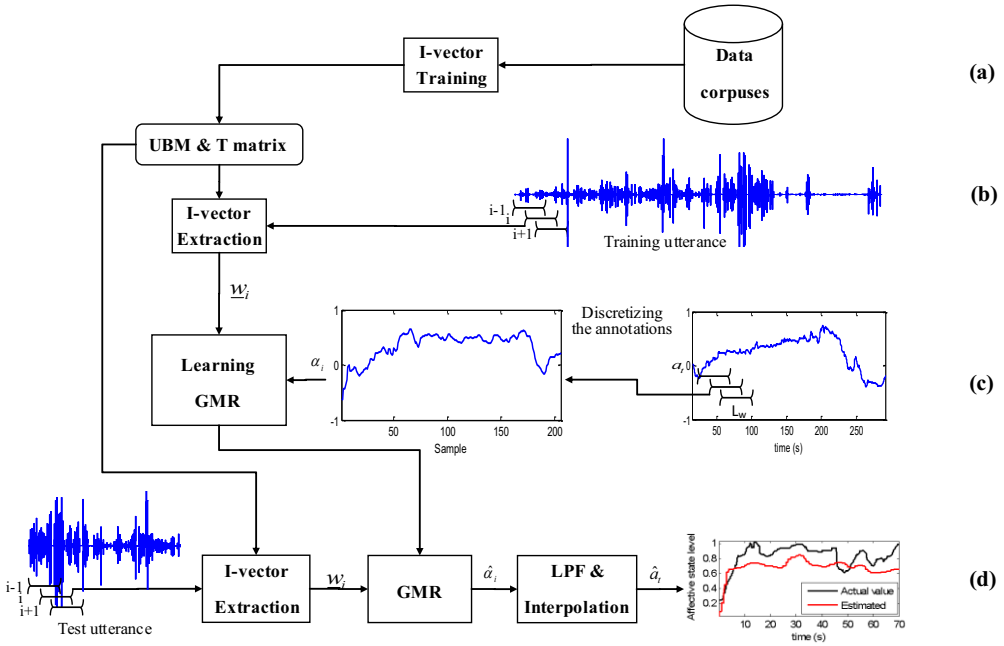

Figure 1: *An overview of the method, (a) TVS training, (b) i-vector extraction, (c) discretizing the features and GMR training, and (d) affective state estimation and low pass filter smoothing.*

each session there are two actors. They improvise an interaction (such as, accepting or rejecting a request) and their speech is recorded by two microphones separately. Each session contains 5 conversations, where each conversation has an average duration of 4-5 minutes. Continuous annotations in AVD space are provided by three or four annotators. The annotation streams are normalized between -1 and 1 after a post-processing of low-pass filtering and delay compensation. The mean of the annotator curves is defined as the affective dimension a_t. Furtheremore the mean of the pairwise correlations of the multiple annotator curves are defined as the ground truth correlation score.

Table 1. *Notations and descriptions of the test conditions*

Estimate	Test name	Test details
$\hat{a}_{t,F}$	Frame level	MFCCs + GMR per frame
$\hat{a}_{t,S}$	Stat_PCA	Statistical functionals of MFCCs + PCA + GMR per window + interpolation
$\hat{a}_{t,M}$	MFCC_PCA	Concatenation of MFCCs over a window + PCA + GMR per window + interpolation
$\hat{a}_{t,I}$	i-vector	i-vector generation over a window + GMR per window + interpolation

3.2. System Setup

We perform the CET in frame and window level resolution. Since the valence is more related to the facial expressions and speech provides poor valence information, we just report tracking results of the activation and dominance dimensions in this study. For both tasks we employ an automatic voice activity detector (VAD) [15] to remove the silent segments of the speech recordings. To obtain same size of voiced data after VAD for both resolutions, we apply VAD on the window level resolution. We keep the windows with more than 40% active speech. For the frame level system, we used the 16.66 msec frame size with 50% overlapped, which are compatible with the annotation rate. We use the α_i as the mean attribute of each window. Figure 2 plots the mean discretization distortion e_d as a function of window duration with fixed time shift equal to 1 sec. We choose the window length as 3 sec, where the distortion is less than 5% for all dimensions.

An acoustic feature vector is computed over each frame. Each acoustic feature is a 39 dimensional vector, which includes the energy, the first 12 MFCCs plus the first and second time derivatives.

We define four different tests. Three of them are in the window level resolution and one of them is in the frame level. The details are given in Table 1. To map the feature space on affective space we used dynamic GMR with 4 mixture components as provided in [4].

The number of Gaussian mixtures for TVS system (M), the dimension of i-vectors and PCA output dimension are set in an eight-fold cross validation to maximize the mean correlation scores. Each fold contains one of the sessions, so that train and test have different actresses/actors. The estimated and annotators' emotion dimensions are low-pass filtered for smoothing. We use 32 Gaussian mixtures with diagonal covariance for TVS and 10 expectation-maximization iterations for the extraction of UBM and T matrix. We employ 30 dimensional i-vectors. The MSR Identity Toolbox [16] is used for UBM and TVS calculation. We adjust the PCA output dimension as 30 for MFCC_PCA and 40 for Stat_PCA tasks.

3.3. Results and Discussion

To evaluate the performance we calculate the mean correlation across recordings between the affective dimension, a_t, and the estimated affective dimension $\hat{a}_{t,I}$ of the i-vector task as,

$$r_I = \frac{1}{K}\sum_{k=1}^{K} \frac{\mathrm{cov}(a_t^k, \hat{a}_{t,I}^k)}{\mathrm{std}(a_t^k)\mathrm{std}(\hat{a}_{t,I}^k)}, \quad (3)$$

where a_I^k represents the affective dimension for recording k, and K is the total number of recordings in the dataset. The mean correlations for other tasks can be defined similarly to be r_F, r_S and r_M. We present the mean correlation in Table 2 for different tasks. The last row of this table is the mean of the ground truth correlation scores calculated similar to [4].

Table 2. *Mean correlation and mean ground truth correlation scores*

Methods	Activation	Dominance
i-vector	0.4737	0.1439
Stat_PCA	0.4784	0.1331
MFCC_PCA	0.4428	0.1300
Frame level	0.2909	0.0956
Mean GT Score	0.6199	0.6200

We observe significant improvement from the frame level to the window level resolution as also observed in [4]. This could be due to the slowly varying nature of the affective state. Among the window level tasks, the MFCC_PCA has the poorest performance. A possible reason is that the principal directions of the concatenated MFCCs are more informative for the speaker and lexicon variability and is not good for emotion recognition. On the other hand, i-vector and Stat_PCA have similar mean correlation performance, especially for the activation dimension.

We would like to check whether the i-vector and statistical features deliver similar correlation performances along with the ground truth correlation scores. To illustrate this relationship we plot correlation score differences of tasks as a function of the underlying ground truth correlation score range in Figure 3. We use intervals of 0.05 ground truth correlation scores. For example, at 0.5 ground truth correlation score we present three mean correlation differences, $r_I - r_F$, $r_I - r_M$ and $r_I - r_S$. These mean correlation differences are calculated from the segments with ground truth correlation scores between 0.5 and 0.55. Note that for this segment i-vector task mean correlation scores are higher than the frame level, MFCC_PCA and Stat_PCA scores. In general i-vector performance is better than the MFCC_PCA and frame level tasks. Also note that i-vector and Stat_PCA do not share similar correlation scores along different ground truth correlation score segments. So one should expect performance benefit from decision fusion of i-vector and Stat_PCA tasks. Hence we apply a decision fusion based on the likelihood calculated in GMR for any time index. We apply hard and soft decision fusions. In the soft fusion we simply find the weighted mean of the estimated affective dimensions, where weights are choosen as the normalized likelihoods. In the hard fusion we choose the estimated affective dimension with the highest likelihood.

Table 3 presents the results of soft and hard fusion between different tasks. For activation, soft fusion works better and the best one comes from the i-vector and Stat_PCA that makes a 2% improvement over unimodal performance. When

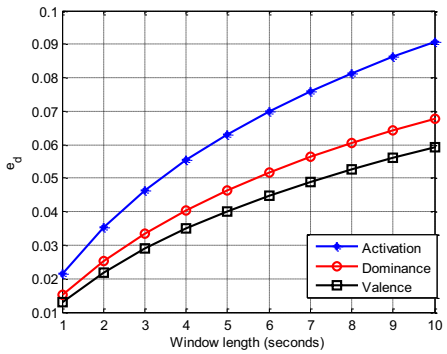

Figure 2: *The discretization distortion versus window length with fixed time shift equal to 1 sec for AVD dimensions using the mean function for discretization*

MFCC_PCA is in the fusion we do not observe significant improvement for activation and dominance estimation. Hard fusion is observed to perform better for dominance estimation for the fusion of i-vector and Stat_PCA (around 2% improvement). Since the dominance estimation has low mean correlation, soft fusion has potential to carry noisy likelihood values to the decision.

Table 3. *Mean correlation of the soft and hard fusion*

Methods	Activation		Dominance	
	Hard	Soft	Hard	Soft
i-vector + Stat_PCA	0.4468	**0.4919**	**0.1639**	0.1452
i-vector + MFCC_PCA	0.4465	0.4757	0.1533	0.1423
MFCC_PCA + Stat_PCA	0.4454	0.4713	0.1339	0.1347
i-vector + Stat_PCA + MFCC_PCA	0.4346	0.4851	0.1574	0.1427

4. Conclusion and Future Work

In this paper, we employ TVS as an unsupervised emotional vocal channel feature extraction framework for continuous tracking of the affective dimensions. Furthermore, we compared it to the other summarization functions. Our experiments were performed on the CreativeIT dataset. The proposed i-vector and Stat_PCA estimators performed similarly in terms of mean correlation score. But we observed that those two estimators are not strongly correlated, hence a decision fusion of i-vector and Stat_PCA estimators attains more than 2% improvement.

In the TVS literature, one of the popular ways of improvement is applying channel compensation techniques. When we apply Within Class Covariance Normalization (WCCN) and Linear Discriminant Analysis (LDA), we do not observe any improvement in the results. We can list at least two reasons for this observation; one is related to the limited number of data for different classes, and the other reason is good channels. Compensation is useful when we have channel variability; however, all of our data were recorded over the microphone and have high quality. Hence, channel compensation methods can be applied to wild data as future work.

5. Acknowledgements

This work was supported by TÜBİTAK under Grant Number 113E102.

6. References

[1] A. Metallinou, Multimodality, context and continuous dynamics for recognition and analysis of emotional states, and applications in healthcare: University of Southern California, 2013.

[2] H. Gunes and M. Pantic, "Automatic, dimensional and continuous emotion recognition," International Journal of Synthetic Emotions (IJSE), vol. 1, pp. 68-99, 2010.

[3] M. K. Greenwald, Cook, E. W., & Lang, P. J. 1989, "Affective judgment and psychophysiological response: Dimensional covariation in the evaluation of pictorial stimuli," Journal of psychophysiology, 1989.

[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] R. W. Levenson, "Emotion and the autonomic nervous system: A prospectus for research on autonomic specificity," Social psychophysiology and emotion: Theory and clinical applications, pp. 17-42, 1988.

[6] Y. Song, L.-P. Morency, and R. Davis, "Action recognition by hierarchical sequence summarization," in Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on, 2013, pp. 3562-3569.

[7] M. H. Bahari, McLaren, M., & Van Leeuwen, D. 2012., "Age estimation from telephone speech using i-vectors," Proceedings Interspeech, pp. 506-509, 2012.

[8] R. Xia, & Liu, Y. 2012., "Using i-Vector Space Model for Emotion Recognition," INTERSPEECH, September 2012.

[9] N. Dehak, Kenny, P., Dehak, R., Dumouchel, P., and Ouellet, P. 2011., "Front-end factor analysis for speaker verification," Audio, Speech, and Language Processing, IEEE Transactions on, pp. 788-798, 2011.

[10] C. Busso, Bulut, M., Lee, C. C., Kazemzadeh, A., Mower, E., Kim, S., and Narayanan, S. S. 2008., "IEMOCAP: Interactive emotional dyadic motion capture database," Language resources and evaluation, pp. 335-359.

[11] M. Grimm, Kroschel, K., & Narayanan, S. 2008., "The Vera am Mittag German audio-visual emotional speech database.," In Multimedia and Expo, 2008 IEEE International Conference on, pp. 865-868, June 2008.

[12] B. Schuller, Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., Müller, C., & Narayanan, S. 2013., "Paralinguistics in speech and language—state-of-the-art and the challenge," Computer Speech & Language, pp. 4-39, 2013.

[13] B. Schuller, Steidl, S., & Batliner, A. 2009, "The INTERSPEECH 2009 emotion challenge," INTERSPEECH, pp. 312-315, September 2009.

[14] A. Metallinou, Lee, C. C., Busso, C., Carnicke, S., Narayanan, S., and Tx, D. 2010, "The USC CreativeIT database: a multimodal database of theatrical improvisation," Workshop on Multimodal Corpora, LREC., May 2010.

[15] M. Brookes. (1997) Voicebox: Speech processing toolbox for matlab. Software. available [Mar. 2011] from www. ee. ic. ac. uk/hp/staff/dmb/voicebox/voicebox. html.

[16] S. M. Sadjadi, Slaney, M., Heck, L. 2013, "MSR Identity Toolbox v1.0: A MATLAB Toolbox for Speaker-Recognition Research," Speech and Language Processing Technical Committee Newsletter, IEEE, November 2013.