Emotion Recognition using Linear Transformations in Combination with Video

Rok Gajšek, Vitomir Štruc, Simon Dobrišek, France Mihelič

LUKS, Faculty of Electrical Engineering
University of Ljubljana, Slovenia
{rok.gajsek, vitomir.struc, simon.dobrisek, france.mihelic}@fe.uni-lj.si

Abstract

The paper discuses the usage of linear transformations of Hidden Markov Models, normally employed for speaker and environment adaptation, as a way of extracting the emotional components from the speech. A constrained version of Maximum Likelihood Linear Regression (CMLLR) transformation is used as a feature for classification of normal or aroused emotional state. We present a procedure of incrementally building a set of speaker independent acoustic models, that are used to estimate the CMLLR transformations for emotion classification. An audio-video database of spontaneous emotions (AvID) is briefly presented since it forms the basis for the evaluation of the proposed method. Emotion classification using the video part of the database is also described and the added value of combining the visual information with the audio features is shown.
Index Terms: emotion recognition, emotional database, linear transformations

1. Introduction

Analysis of the psycho physical state of participants in a modern audio/video (human-human or human-machine) communications is becoming an important field of research. Both, in speech recognition and face detection communities, there is an increasing aspiration for improving emotion detection and thus, providing additional information to the system. In the case of human-human communications the benefit of emotion or arousal detection can assist in a more secure and credible exchange of information, whereas in human-machine interaction advantages can be gained in speaker identification/verification, development of dialog managers, speech recognition, etc.

Optimistic recognition rates have been obtained when using acted emotional databases [1, 2, 3], but the scores drop significantly if spontaneous emotions are being recognised. The effect is similar than in the case of recognising spontaneous speech as oppose to the read or broadcast news type of speech. Spontaneous emotions, recorded from people in real life when they interact with a system or another person, tend to be less explicit then the actor's attempt to represent a given emotion. Also, one can argue that the acted emotions do not adequately represent the real life emotions people experience every day. In addition, the emotion labelling in the case of spontaneous emotions presents a greater challenge since there is no prior knowledge on what emotion or arousal state is being represented. A brief description of the AvID [4], audio-video emotional database of spontaneous emotions, is given, as it forms the basis for evaluation of the proposed methods.

The work, presented in this paper, focuses on recognising emotions based on the linear transformations of the Hidden Markov Models (HMM), usually used in speaker or environment adaptation. The constrained maximum likelihood linear regression technique (CMLLR) [5] is evaluated, first as a standalone feature for emotion recognition, and also in combination with the video. For completeness, the video emotion recognition sub-system is also presented.

2. AvID: Audio-video database of spontaneous emotions

AvID database is comprised from two recording sessions. In the first session, called "Adaptive IQ test" the participants were deceived into thinking that the aim of the experiment is to examine how different biometric measures can be used in an adaptive test of intelligence. First, they had to describe photographs with neutral content, supposedly to measure his/her verbal fluency, but the real goal was to acquire recordings in normal, non-emotional state. Secondly, they played a game of Tetris, but instead of using the keyboard as input, they had to lead the experimenter through the game by giving verbal instructions. Permitted commands were slovenian translations of left, right, down and around. The explanation, given to the participants for playing the game, was that efficiency of his/her verbal instruction given to a teammate to achieved a common goal, will be assessed. The game was recorded and, in the third part of the session, played back to the participant who now had to describe what is happening on the screen by uttering the same four commands. The final part of this recording session was the adaptive IQ test, comprised of twenty 3 by 3 matrices with one element missing. The participant had to reason aloud about the principals of the agreement of matrix elements and had to find the logical solution among five or six proposed answers. Since the aim of the test was to induce spontaneous arousal, the current IQ estimate, participant's hear beat and time left to finish the task, were displayed on the screen. With first few, simple matrices the IQ value increased slightly, but as the test progressed the IQ value began dropping regardless of the answers. Also the displayed heart beat was not of the actual participant but was manipulated by psychologist so it kept increasing throughout the test. At the end of session the participants were debriefed and the experimentator explained the real goal of the study.

As opposed to the first part where different levels of arousal were induced, the second part targeted the major six emotions: happiness, anger, surprise, disgust, fear and sadness. The participant watched a short video (approximately 10 minutes) aiming at inducing a particular emotion. After the video, they were presented with a set of photographs targeting the same emotion. He/she had to present their thoughts on the observed material, how they feel about it, if it relates to anything in their life, etc.

Sessions were recorded using a high-definition camera and three different microphones to enable some environment and channel normalisation tests. One microphone was a clip-on microphone and was positioned on the chest of the participant, the second was placed on the nearby table and for the third channel the camera's microphone was used. A total of 19 participants were recorded in the adaptive IQ test (12 female and 7 male) and 9 in emotional videos session so far. All together, that comprises to approximately 30 hours of recordings. The process of transcribing and labelling the data according to the emotion is currently under way.

3. The audio sub-system

Generally, in emotion recognition from speech the emphasis is on analysis of prosody. Therefore, a large number of prosody based features exist ranging from prosodic features extracted directly from audio signal (pitch, energy, voicing, etc) or derived prosodic features which are computed using speech recogniser (phone-duration, speaking rate, etc.) [6]. Combined with a set of cepstrum based coefficients they form a feature set used for classification.

3.1. Maximum likelihood linear regression

Linear transformations are an effective way of adapting the HMMs to a different conditions, either different speaker or different environment. In general the transformations can be calculated in different ways, but if maximum likelihood estimation is used the transforms are called MLLR. These transformation matrices are trained on a new data set in order to maximises the likelihood of the new data set given the original model. Generally, only the means and variances of the HMMs are transformed and the transition probability distribution and a-priori probabilities are left unchanged. Two different transformations can be used: an unconstrained, meaning that there are a two different transformations for mean and variance, or constrained, where the same transformation matrix is used for both means and variances.

In our work we focused on using a constrained version (CMLLR) [5] that can be easily converted from model space (transformation of means and variances) to feature space (transformation of acoustic features). The vector of means μ and the covariance matrices Σ are transformed in the case of MLLR according to the following equations

$$\hat{\mu} = \mathbf{A}'\mu - \mathbf{b}', \tag{1}$$

$$\hat{\Sigma} = \mathbf{A}'\mathbf{\Sigma}\mathbf{A}'^{\mathbf{T}}. \tag{2}$$

The matrix $\mathbf{A}'$ and the vector $\mathbf{b}'$ form the linear transformation and their parameters are determined by maximising the likelihood of the acoustic model based on the feature vector set $\mathbf{o}(\tau)$, which is available for adaptation. A well known Expectation-Maximisation (EM) algorithm [5] is used to estimate the coefficients of $\mathbf{A}'$ and $\mathbf{b}'$.

The above described procedure yields a state space transformation which is used to transform the parameters of Gaussian distributions. But $\mathbf{A}'$ and $\mathbf{b}'$ can also be used in a feature space, meaning that the acoustic features can be modified, instead of the model itself. This is presented by the following equation

$$\hat{\mathbf{o}(\tau)} = \mathbf{A}'^{-\mathbf{1}}\mathbf{o}(\tau) + \mathbf{A}'^{-\mathbf{1}}\mathbf{b}' = \mathbf{A}\mathbf{o}(\tau) + \mathbf{b}. \tag{3}$$

The matrix $\mathbf{A}$ and the vector $\mathbf{b}$ are usually combined into a single matrix $\mathbf{W} = [\mathbf{A}\ \ \mathbf{b}]$ which represent the full transformation.

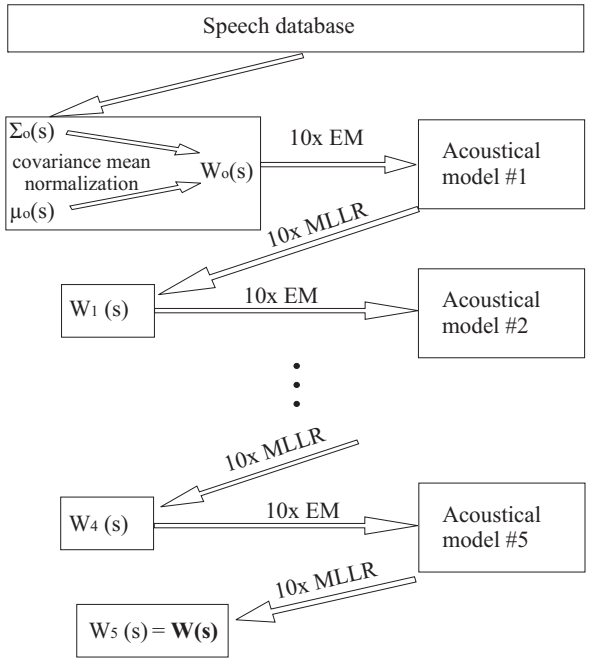

Figure 1: *Acoustic model training procedure*

3.2. Acoustic model training

In speaker adaptation the aim is to "move" the speaker specific characteristics from the acoustic model to the MLLR transform and thus making the acoustical model speaker independent. We used the same tactics, except instead of moving the speaker specific elements to MLLRs we aimed at shifting the emotional information from the acoustic model to the transform matrix. Therefore the speaker dependant information should be removed from the acoustic model before estimation of the emotional MLLRs. Hence, we devised a training procedure shown in Fig. 1 in order to first train a speaker independent acoustic model.

The database used for building the acoustic models was Voicetran [7]. First the starting matrix $\mathbf{W_0}$ is initialised by normalisation of means and covariance matrix. This is accomplished using equations

$$\mathbf{A(s)} = \mathbf{L_0(s)}^{\mathbf{T}}, \mathbf{b(s)} = -\mathbf{L_0(s)}^{\mathbf{T}}\mu_{\mathbf{o}}(\mathbf{s}). \tag{4}$$

where $\mu_o(s)$ is the global vector of means and $\mathbf{L_0(s)}$ is a Cholesky decomposition matrix of the inverse of the global covariance matrix $\mathbf{\Sigma_0(s)}^{-\mathbf{1}}$ and s represents the speaker. The first transformation matrix $\mathbf{W_0} = [\mathbf{A(s)}\ \ \mathbf{b(s)}]$ is used for 10 passes of training a monophone acoustic model #1. Using the model #1 the new MLLR transformation for each speaker $\mathbf{W_1(s)}$ can be calculated, again by 10 passes of EM. The $\mathbf{W_1(s)}$ MLLRs are then used to train the acoustic model #2. We repeat the process through five iterations and end up with the final set of $\mathbf{W_5(s)}$ MLLRs which are used for speaker adaptation. The reason for applying the above procedure is to obtain five acoustic models needed for estimation of the set of MLLRs that will carry emotion specific information.

3.3. MLLR estimation for emotion recognition

For emotion recognition, the AvID emotional database was used. Six sessions, both transcribed and labelled as emotionally neutral or aroused, were taken to form the bases for the evaluation tasks. The sentence-based utterances were grouped into 15 seconds long parts according to the speaker and emotion label. This is needed to provide enough data to estimate a MLLR matrix. The above described process of five iterations of estimating the MLLRs, without the acoustic model re-estimation (the acoustic models are used unmodified), is applied to generate the final set of MLLRs which are used for emotion recognition.

4. The video sub-system

Recognising an individuals emotional state from video data requires a number of tasks to be performed. These tasks commonly include: *(i)* the detection of the facial region in each of the video-frames, *(ii)* the geometric and photometric normalisation of the detected facial region, *(iii)* the extraction of suitable visual features from the normalised facial region of each video frame, and *(iv)* the classification of the given video sequence into one of the classes corresponding to the emotional states.

Each of the listed tasks contributes to the overall efficiency of the visual part of the (emotion) recognition system and, therefore, needs to be solved properly. Since the output of the first task represents the input for second one, the output of the second task represents the input for the third one and so forth, errors and deficiencies at each of the tasks are likely to be propagated throughout the system, influencing the systems final recognition performance. Clearly, a robust solution for all tasks in the visual processing chain is needed to devise an effective emotion recognition system. Our solution to each of the steps comprising the visual processing chain are presented in the remainder of this section.

4.1. Face detection

The first step of each recognition system relying on visual features represents the extraction of the so-called region of interest (ROI) from the input image. In case of an (visual) emotion recognition system, where the goal is to classify video sequences of facial images into classes of emotional states, this translates into the detection of the facial region in each of the frames of the video sequence. This step removes all artifacts not belonging to the face from the video frames and, hence, ensures that the extracted visual features encode only information relevant to the recognition stage.

Several techniques have been presented in the literature to perform face detection. However, due to its efficiency and real-time capabilities we make use of the popular Viola-Jones face detector (VJFD) [8]: the key component of the VJFD is an image representation called *integral image*, which allows visual features of image sub-windows of arbitrary sizes to be computed in constant time. Once the features over a predefined number of sub-window sizes have been computed, AdaBoost is exploited to select a small set of important visual features that are ultimately fed to a cascade classifier. The classifier then assigns each sub-window of the processed image either to the class of "faces" or the class of "non-faces".

It has to be noted that the VJFD, when properly trained, usually results in a satisfactory face-detection-performance; however, a problem of the VJFD are the false positives, i.e., image sub-windows falsely assigned to the class of "faces". To compensate for this deficiency and consequently improve the face detectors performance, we further process the detectors outputs using a skin-colour filter.

4.2. Face window normalization

After the detection of the facial region in a given video frame, the image sub-window corresponding to the face needs to be normalised. Faces can generally be tilted in one direction and can exhibit illumination induced appearance variations. The normalisation procedure tries to compensate for these irregularities by rotating the face by an angle which ensures that the line passing through both eyes is in a horizontal position and by subjecting the rotated image to a histogram equalisation procedure which enhances the images contrast and ensures invariance to (moderate) illumination variations. Finally, the normalised face region is re-scaled to a standard size of 128×128 pixels. An example of the adopted normalisation procedure is shown in Fig. 2.

Figure 2: An example of the normalisation procedure (from left to right): the output of the VJFD, the aligned facial region, and the histogram equalised and re-scaled facial image

While histogram equalisation is a common preprocessing step in image processing and, hence, needs no further explanation, determining the rotation angle is a non-trivial task and deserves a more detailed description. A simple, but effective procedure was proposed by Pozne in [9]. The author suggested to use the integral projection of the horizontal image gradient onto the vertical axes as a criterion for the selection of the rotation angle. Here, the image is rotated by a number of angles and at each of these angles the integral projection onto the vertical axes is computed. Since the facial landmarks occur in pairs (e.g., eyes, eyebrows, etc.), an aligned face is expected to produce higher peaks and a characteristic shape when projected onto the vertical axes. Based on the integral projection of the image the best rotation angle is determined and employed to align the face. Due to its efficiency the presented procedure was also adopted in our work.

4.3. Visual feature extraction

The normalised facial regions extracted from each frame of the video sequence form the foundation for the feature extraction procedure. Here, the 2-dimensional discrete cosine transform (DCT) is employed as the feature extraction technique. The DCT is a popular image processing tool commonly used for image compression. When adopted as a feature extraction approach it is used to reduce the dimensionality of input frames by retaining only the first[1] d DCT coefficients, where $d << N$ and N stands for the number of pixels in the normalised facial region, i.e., $N = n \times n = 128 \times 128$. In our case, the value of d was set to 300.

Formally, the 1-dimensional DCT transform on a n-dimensional sequence $u(i)$, where $i = 1, 2, ..., n$, is defined

[1] Note that the term "first" refers to the first coefficients of the DCT transformed image sampled in the classical zig-zag manner.

as follows:

$$v(k) = \alpha(k) \sum_{i=0}^{n-1} u(i) cos(\frac{(2i+1)\pi k}{2n}), \quad (5)$$

where

$$\alpha(0) = \sqrt{\frac{1}{n}}, \text{ and } \alpha(k) = \sqrt{\frac{2}{n}}, \text{ for } 1 \leq k \leq n-1. \quad (6)$$

In the above expressions $v(k)$ denotes the DCT transformed sequence $u(i)$. Since the DCT transform represents a separable transform its 2-dimensional variant is obtained by first applying the 1-dimensional DCT to all image rows (which act as the sequences $u(i)$) and then to all image columns.

5. Training and classification

The work presented in this paper represents our initial attempts in devising an audio-visual emotion recognition system. A simple nearest neighbour classifier was therefore selected for the classification phase. As the goal of the system is to classify the audio-video sequence into one of two possible classes, the class prototypes (or models) for audio and video need to be built during the training session. Prior to building the models, a feature vector was constructed by combining all the columns of each 39 by 39 MLLR matrix. A full set of 668 vectors (629 representing normal state and 39 aroused state) was divided into training (80%) and testing data (20%). Next, the vectors were transformed by using principal component analysis (PCA) to a smaller dimension of 531 (this is the maximum possible size, based on the number of train cases). From the training data, two models in form of the mean feature vectors are constructed, one for the neutral emotional state and one for aroused emotional state. The means are taken over all subjects, regardless of the identity. During the classification procedure, each MLLR derived vector from the testing part is ultimately classified based on the ratio between the Euclidean distance of the feature vector and the model of the neutral emotional state, and the distance between the feature vector and the aroused emotional state. Formally, this can be written as:

$$\frac{d_n(\mathbf{x}, \boldsymbol{\mu}_n)}{d_a(\mathbf{x}, \boldsymbol{\mu}_a)} > \Delta, \quad (7)$$

where $\mathbf{x}$ denotes the feature vector sequence (i.e, matrix of feature vectors), $\boldsymbol{\mu}_n$ denotes the mean feature vector of the neutral emotional state, $\boldsymbol{\mu}_a$ stands for mean feature vector of the aroused emotional state, d_i ($i \in \{n, a\}$) stands for the function returning the Euclidean distance, and Δ represents a predefined decision threshold. In case the ratio is above the selected decision threshold Δ the video sequence is assigned to the class of the neutral emotional states, otherwise it is classified as belonging to the class of aroused emotional states.

A similar approach was used for the video part where, instead of using PCA transformed MLLR matrices, visual features, as described in Section 4.3, were used. Distance ratios (7) from both, the audio and video classification are combined using sum rule fusion.

6. Results

A 5-fold cross validation was used in order to obtain a more objective results. Each time 80% of the data was used for training and the remaining 20% for test. An average of all 5 cross validation runs are shown in Table 1.

Table 1: *Percentage of correctly classified emotional states with and without the video.*

	Neutral	Aroused	Weighted
Audio only	66.2%	64.3%	**66.0%**
Audio + Video	71.2%	72.5%	**71.3%**

First, the results of emotion recognition using just the CMLLR transformations as features are shown. The first and second columns shows the correctly classified Neutral and Aroused states, respectively, the second column is the combined score, weighted according to the number of test cases for each class. It can be seen that accuracy for both, neutral and aroused classes, is around the weighted value of 66%. In the second row, the combined results for audio and video are shown. The jump for neutral class, from using only audio, is 5%, but interestingly, the increase in recognition rate is almost 8%.

7. Conclusion

In the paper, we presented our initial attempts on building an audio-video emotion recognition system. The idea of employing linear transformations, otherwise heavily used for speaker or environment adaptation, as classification features for emotion recognition was presented. The usage of one type of linear transformation, called CMLLR, is evaluated on the AvID database of spontaneous emotions. We have shown that using only CMLLR transformations and a simple classification method can give around 66% of correctly recognised emotion states. Furthermore, with inclusion of the video information the recognition rate roused to 71.3%. The future work lies in evaluating the described procedure on the full AvID emotional database, once it is fully labelled, and also test other emotional databases.

8. References

[1] Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W., Weiss, A., "A Database of German Emotional Speech", Proceedings Interspeech 2005: 1–4., 2005.

[2] LDC, "SUSAS (Speech Under Simulated and Actual Stress)", Proceedings of the 4th AFGR'00, Language Data Consortium, 1999.

[3] LDC "Emotional Prosody Speech and Transcripts", Language Data Consortium, 2002.

[4] Gajšek, R., et al, "Multi-Modal Emotional Database: AvID", Informatica 33: 101–106, 2009

[5] Gales, M. J. F., "Maximum likelihood linear transformations for hmm-based speech recognition", Computer Speech and Language, 12 (2): 75–98, 1998.

[6] Shriberg, E. and Stolcke, A., "Direct Modeling of Prosody: An Overview of Applications in Automatic Speech Processing", Proc. International Conference on Speech Prosody, 2004.

[7] Mihelič, F., et al, "Spoken language resources at LUKS of the University of Ljubljana", Int. J. of Speech Technology, 6 (3): 221–232, 2006.

[8] Viola, P. and Jones M., "Robust real-time object detection", in Proc. of the Second International Workshop on Statistical and Computational Theories of Vision - Modeling, Learning, Computing and Sampling, Vancouver, Canada, July 2001.

[9] Pozne, A., "Extracting visual features for automatic speech recognition", PhD thesis, Faculty of Electrical Engineering, University of Ljubljana, 2005.