A DIMENSIONAL APPROACH TO EMOTION RECOGNITION OF SPEECH FROM MOVIES

Theodoros Giannakopoulos, Aggelos Pikrakis and Sergios Theodoridis

Dept. of Informatics and Telecommunications
University of Athens, Greece
e-mail: {tyiannak, pikrakis, stheodor}@di.uoa.gr
URL: http://www.di.uoa.gr/dsp

ABSTRACT

In this paper we present a novel method for extracting affective information from movies, based on speech data. The method is based on a 2-D representation of speech emotions (Emotion Wheel). The goal is twofold. First, to investigate whether the Emotion Wheel offers a good representation for emotions associated with speech signals. To this end, several humans have manually annotated speech data from movies using the Emotion Wheel and the level of disagreement has been computed as a measure of representation quality. The results indicate that the emotion wheel is a good representation of emotions in speech data. Second, a regression approach is adopted, in order to predict the location of an unknown speech segment in the Emotion Wheel. Each speech segment is represented by a vector of ten audio features. The results indicate that the resulting architecture can estimate emotion states of speech from movies, with sufficient accuracy.

Index Terms— Multimedia analysis, Emotion Recognition, Regression

1. INTRODUCTION

The enormous increase in storing multimedia content has led to research efforts for content-based information analysis on the audio-visual signals. Apart from extracting information regarding events, structures (e.g., scenes) or genres, a substantial research effort has focused on recognizing the affective content of multimedia material, i.e., the emotions that underlie the audio-visual information ([1], [2], [6]). Automatic recognition of emotions in multimedia content can be very important for various multimedia applications. For example, recognizing affective content of music signals ([3], [4]) can be used in a system, where the users will be able to retrieve musical data with regard to affective content. In a similar way, affective content recognition in video data could be used for retrieving videos that contain specific emotions. In this paper, we emphasize on affective content that can be retrieved from the speech information of movies.

The most common approach to affective audio content recognition, so far, is to apply well-known classifiers (Hidden Markov Models, etc.) for classifying signals into discrete categories of emotions, e.g., fear ([1], [7]). One drawback of such techniques is that, in many cases, the emotions of multimedia content cannot easily be classified in specific categories. For example, a speech segment from a horror movie may contain both fear and disgust feelings. In addition, the level of categorical taxonomy of emotion is subjective, i.e., the number of classes is an ambiguous subject. For example, the state of happiness can be further divided into pleasure and excitement.

An alternative way to emotion analysis is the dimensional approach, according to which, emotions can be represented using specific dimensions that stem from psychophysiology ([5], [4]). In [5], Valence-Arousal representation is used for affective video characterization. Towards this end, visual cues, such as motion activity, and simple audio features, e.g., signal energy are used for modelling the emotion dimensions. In [4], the same representation is used for estimating music emotions using regression techniques.

In this paper, the problem of speech emotion recognition is also treated as a regression task: 10 audio features are mapped to the Valence and Arousal dimensions using the kNN method. To our knowledge, this is the first time that the dimensional approach is studied explicitly for speech emotion recognition from real video content. The contribution of this work is focused on the following:

1. To investigate whether the dimensional representation of Arousal-Valence is appropriate for speech signals. Towards this end, several humans have manually annotated speech segments using this representation. If the Emotion Wheel is a good representation, then the differentiation by separate humans should be, on average, in good agreement.

2. An extensive research has lead to the selection of some audio features for the specific regression problem.

3. The proposed regression scheme is evaluated using the annotated data, and the performance error is compared to the error of the human annotation.

2. REPRESENTATION AND DATA COLLECTION

2.1. 2-D Emotional Representation

Dimensional emotion representation ([5], [6]) is based on some psychological understandings. In particular, the emotion plane is viewed as a continuous 2-D space where each point corresponds to a separate emotion state. The two dimensions of this plane are valence (V) and arousal (A). Valence varies from -1 (unpleasant) to 1 (pleasant) and therefore it can be characterized as the level of pleasure. Arousal, on the other hand, represents the intensity of the affective state and it ranges from -1 (passive, calm) to 1 (active). Each emotional state can be understood as a linear combination of these two dimensions. Anger, for example, can be conceptualized as an unpleasant emotional state (negative V-values) with high intensity (positive A-values). In Figure 1 a scheme of the 2-D emotional representation is presented (usually called "Emotion Wheel" - EW), along with some basic emotional states and their (approximate) positions in the plane.

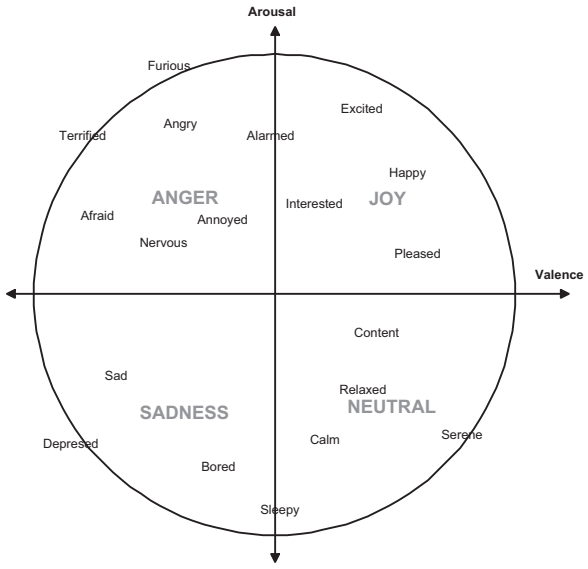

Fig. 1. 2-Dimensional Affective Representation

2.2. Emotional Data Collection

In order to evaluate the representation presented in Section 2.1 and also to train and test the proposed emotional recognition method, we have manually selected 1500 speech audio samples from more than 30 films. The films were selected to cover a wide range of genres (e.g. horror, comedy, etc). The average duration of the segments is 2.5 seconds. The manual annotation of speech emotion was accomplished by 30 humans. In particular, each human randomly listened to a number of speech segments. For each speech segment he/she selected a point in the emotion plane, according to the estimated emotion. It has to be emphasized that, each time, the users were prompt with a random audio sample and the same sample could appear in a later annotation. In this way, we could compute the level of disagreement among annotations, of the same segment, by the same user. The manually annotated data was therefore used for three purposes, namely:

1. Train (and test) the proposed automatic emotion recognition method. Towards this end, for each audio sample i, if the number of annotations N_i was larger than 5 (i.e., at least 5 humans have annotated this sample), the average annotated coordinates were used as the final coordinates. In other words, for each sample i, with user-annotated coordinates: $xs_{ij}, i = 1, \ldots, N_i$ and $ys_{ij}, i = 1, \ldots, N_i$, the final emotion coordinates were $x_i = \frac{\sum_{j=1}^{N_i} xs_{ij}}{N_i}$ and $y_i = \frac{\sum_{j=1}^{N_i} ys_{ij}}{N_i}$.
2. Evaluate the level of disagreement among the different users. Suppose that A_j (length L_j) is an array that contains the indices of the audio segments that have been annotated at least once by user j, and also have been annotated by at least 5 users in total. We have decided to use the average normalized Euclidian distance of the decisions of this user from the respective average decisions, as a measure of disagreement:

$$D_j = \frac{1}{L_j} \cdot \sum_{i \in A_j} \frac{\sqrt{(xs_{ij} - x_i)^2 + (ys_{ij} - y_i)^2}}{\sqrt{x_i^2 + y_i^2}} \quad (1)$$

3. Evaluate the level of disagreement for the annotation decisions of the same user. Towards this end, we detect the audio segments which have been annotated at least twice by user j. For each one of those audio segments, we calculate the average user decision (i.e. average emotion coordinates) and then the average normalized distance of all decisions from that average value. Finally, DS_j is computed by averaging normalized distances for all audio segments. Therefore, DS_j is a measure of (normalized) deviation of the $j-$th user's annotation decisions. We will refer to this as "self-error".

3. AUTOMATIC SPEECH EMOTION RECOGNITION

3.1. Audio Features

For each audio segment, 10 features and respective statistics are extracted. In particular, a short-term processing is applied: each audio segment is broken into non-overlapping short-term windows (frames) and for each frame a feature value is calculated. Then, for the extracted feature sequence, a statistic is computed (e.g., standard deviation). This statistic is the final feature value that characterizes the whole segment. The following features / statistics have been used ([8], [9]):

1. The average value of the 3rd MFCC.
2. The maximum value of the 2nd MFCC.
3. For each 20 mseconds frame the FFT is computed and the position of the maximum FFT value is kept. Then, the maximum value of that sequence is the final feature for the audio segment.
4. This feature is also based on the position of the maximum FFT bins, though, this time the adopted statistic is the standard deviation of the sequence.
5. The Zero Crossing Rate is firstly calculated on a short-term basis (20 mseconds). The adopted statistic is the standard deviation to average ratio ($\frac{\sigma^2}{\mu}$).
6. The median value of the Zero Crossing Rate sequence.
7. The $\frac{\sigma^2}{\mu}$ ratio of the Spectral Centroid sequence.
8. The $\frac{max}{\mu}$ ratio of the pitch sequence. The pitch was calculated using the autocorrelation method.
9. The $\frac{\sigma^2}{\mu}$ ratio of the pitch sequence.
10. The 2nd chroma-based feature, described in [9], which is a measure of variation of chroma elements over successive short-term frames.

The features and statistics have been selected after extensive experimentation. Furthermore, most of the features have a physical meaning for the specific problem. For example, the $\frac{\sigma^2}{\mu}$ ratio of the pitch sequence shares high values for audio segments generally characterized as "anger", since speech under such an emotional state has large pitch variations.

3.2. Regression

As explained in Section 3.1, for each audio sample a $10-$D feature vector is computed. Furthermore, each speech segment i is represented using two continuous values (x_i, y_i), which express the respective position in the EW. Therefore, we need to train two regression models that map the 10 features to the corresponding emotion dimensions, i.e. two functions $f_1, f_2 : \mathbb{R}^{10} \rightarrow \mathbb{R}$.

We have chosen the k-Nearest Neighbor rule in its regression mode ([10]), since it is a simple and efficient way to estimate the values of an unknown function, given a number of training points. The training data is described by the x and y coordinates: $X = \{x_i\}$ and $Y = \{y_i\}$ and the respective feature vectors $\mathbf{F} = \{\overline{F_i}\}$, $i = 1 \ldots K$, where K is the total number of training samples (i.e., the number of samples that have been annotated by at least 5 humans). Given those training sets and an audio segment described by a $10-$D feature vector F_{test}, we need to estimate the emotion wheel coordinates of that audio segment: x'_i and y'_i. Towards this end, we form the subsets $N_1 \subset X$ and $N_2 \subset Y$, composed by those elements whose respective feature vectors (of **F**) are the k-nearest to F_{test}. The kNN estimation is then applied for both dimensions, according to the following equations:

$$x' = \hat{f}_1(\overline{F}_{test}) = \frac{1}{k} \sum_{x \in N_1} x \quad (2)$$

$$y' = \hat{f}_2(\overline{F}_{test}) = \frac{1}{k} \sum_{y \in N_2} y \quad (3)$$

In order to estimate the emotion coordinates for a set of (test) samples, we compute the following error measure, which is the average distance between the real and estimated coordinates, normalized by the distance of the real coordinates from the $(0, 0)$ point of the EW:

$$E = \frac{1}{K} \cdot \sum_i \frac{\sqrt{(x_i - x'_i)^2 + (y_i - y'_i)^2}}{\sqrt{x'^2_i + y'^2_i}} \quad (4)$$

4. EXPERIMENTS

4.1. Emotion Representation Evaluation

As discussed in Section 2, D_j and DS_j correspond to the $j-$th user's normalized distance from the average decisions and normalized distance from the same user's mean decision. These two quantities are used to evaluate the 2D emotion representation itself. In table 1, the average, minimum and maximum values of these measures are presented. It can be seen that the average error of the users' annotation decisions is quite low (0.85), i.e. it is (on average) equal to 85% of the sample's true position from the center of the EW. This is indicative that the EW offers a good affective representation for speech segments.

Measure	Average	Min	Max
User Error (D)	0.85	0.59	1.18
Self-User Error (DS)	0.53	0.22	1.08

Table 1. Measures of annotation agreement

4.2. Emotion Recognition Evaluation

For testing and training purposes, the K audio samples (i.e., the number of speech segments annotated by at least 5 humans) have been used. After the completion of the annotation procedure, K was equal to 350. For training the kNN regression scheme, 60% of the samples were used, while the remaining samples were used for testing purposes. For the final experiments, cross-validation has been used. In particular, 500 repetitions of random sub-sampling validation have been executed. The average normalized error (described in Equation 4) was found to be 0.97. For comparison purposes,

and in order to have a worst case scenario, we have computed the same error measure for the random estimator of emotion coordinates, i.e. by selecting randomly (x, y) in the EW. For this random scheme, the error was found to be 1.98. In Figure 2 the distribution of D and DS error measures for all humans is presented, along with their average values and the kNN and random estimators' errors. The automatic regression method gives a prediction very close to the average of the respective human decisions. Moreover, the experiments indicate that the kNN regression model achieves better performance than 19% of the humans.

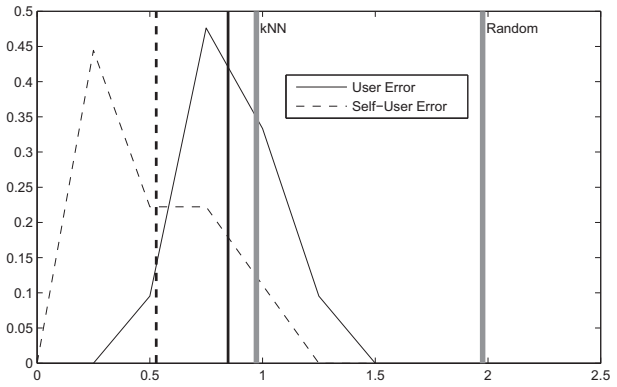

Fig. 2. D and DS error distribution. Vertical lines represent the average values of the error distributions, while the gray vertical lines correspond to the errors of the kNN algorithm (E) and the random estimator.

In addition, in order to check the regression performance in each one of the two dimensions (i.e. Valence - Arousal), the R^2 statistic ([11]) was used. The following values were obtained:

- Valence: 24%
- Arousal: 35%

The previous values indicate that estimating "pleasantness" of an emotion is harder than estimating its intensity.

5. CONCLUSIONS

A dimensional approach to emotion recognition of speech from movies has been proposed. The 2D representation scheme (emotion wheel) has been evaluated using annotations from several humans. The experimental evaluation indicates that the average user error is small, which means that the users' disagreement is also low, and therefore the representation itself can be used in the context of identifying emotions in speech signals. Ten features have been proposed for describing the audio segments and the kNN method has been used for mapping this 10-D feature space to each dimension of the emotion wheel. The normalized error of this regression approach was found to be equal to 0.97, which is very close to the human average annotation error.

6. REFERENCES

[1] Wang, Y. and Guan, L. *Recognizing Human Emotional State From Audiovisual Signals* Multimedia, IEEE Transactions on, 2008, 10 (5), 936-946

[2] Cowie, R.; Douglas-Cowie, E.; Tsapatsoulis, N.; Votsis, G.; Kollias, S.; Fellenz, W. and Taylor, J. *Emotion recognition in human-computer interaction*, Signal Processing Magazine, IEEE, 2001, 18, 32–80

[3] L. Lu and D. Liu and H. Zhang *Automatic mood detection and tracking of music audio signals* , 2006 IEEE Trans. on Audio, Speech & Language Processing, vol 14, 5–18

[4] Yi-Hsuan Yang; Yu-Ching Lin; Ya-Fan Su; Chen, H.H *A Regression Approach to Music Emotion Recognition*, 2008 IEEE Transactions on Audio, Speech & Language Processing, vol 16 (2), pp 448–457

[5] Hanjalic, A. Li-Qun Xu *Affective video content representation and modeling* , 2005 IEEE Transactions on Multimedia, vol 7 (1), pp 143–154

[6] Hanjalic, A. *Extracting moods from pictures and sounds: towards truly personalized TV* , 2006 Signal Processing Magazine, IEEE, vol 23 (2), pp 90–100

[7] A. Nogueiras, A. Moreno, A. Bonafonte, and J. B. Mario *Speech emotion recognition using hidden Markov models*, 2001 in Proc. Eurospeech, pp. 2679-2682

[8] S. Theodoridis and K. Koutroumbas, *Pattern Recognition, 3d edition*. Academic Press, 2005.

[9] T. Giannakopoulos, A. Pikrakis and S. Theodoridis *Music Tracking in Audio Streams from Movies* , 2008 International Workshop on Multimedia Signal Processing, IEEE Signal Processing Society (MMSP2008)

[10] A. Navot, L. Shpigelman, N. Tishby and E. Vaadia *Nearest Neighbor Based Feature Selection for Regression and its Application to Neural Activity* Advances in Neural Information Processing Systems, 2005

[11] Sen, A. and M. Srivastava *Regression analysis, theory, methods and applications*, M. Springer 1990