SPEECH-BASED EMOTION RECOGNITION USING SEQUENCE DISCRIMINANT SUPPORT VECTOR MACHINES

Talieh Seyed Tabatabaei, Sridhar Krishnan
Department of Electrical and Computer Engineering
Ryerson University
Toronto, Canada

Aziz Guergachi
Department of Information Technology Management
Ryerson University
Toronto, Canada

ABSTRACT

Automatic Emotion Recognition (AER) is an interesting and recent research topic in the Human-Computer Interaction (HCI) field. In this paper a speaker- independent Automatic Human Emotion Recognition system is presented which is able to classify six discrete emotional states: happiness, sadness, anger, surprise, fear, and disgust. A set of novel and robust acoustic features are presented which are proved to yield a very good result. Least Square-Support Vector Machines (LS-SVMs) are proposed as a very powerful classifier with many advantages over other popular classifiers. In order to be able to discriminate between the whole sequences rather than frames, the use of Fisher kernels which make use of the information in the underlying generative models is suggested. Fuzzy-pairwise method is implemented to extend the binary SVMs to our multi-class problem. The overall classification rate of 97.65% is achieved.

INTRODUCTION

Computers are becoming more and more popular every day and accordingly the study of interaction between human (users) and computers is catching more attention. In order to have a more natural and friendly interface between human and computers, it would be beneficial to give computers the ability to recognize situations the same way a human does. Being equipped with emotion recognition system, computers will be able to recognize their users' emotional states and show the right reaction to that.

There are different indicators for different emotional states, like some physical changes in body (e.g. blood pressure, heart rate, muscles being tense or relaxed,…), body gesture (highly person-dependent), lexical meaning, speech signal (acoustic information), facial expressions, and etc. Auditory channel is one of the most important communication channels between people. Therefore, we use acoustic information of the speech signal in order to develop our AER system.

While different kinds and number of emotion have been categorized by different researchers [1, 2, 3], the emotion grouping in this work consists of six discrete emotional states, which are common in all cultures: *anger, happiness, fear, sadness, surprise,* and *disgust.*

Choosing the right classifier which works well for one data set is a very important part of the system. There is no established classifier in the literature which is guaranteed to work well with all kinds of data sets. Various different classifiers have been taken into consideration for categorizing the emotional states. The most common classifiers used are Hidden Markov Model (HMM) [5] and Neural Networks (NNs) [6], whereas the number of works which use Support Vector Machines (SVMs) are relatively very few [4]. SVM is a relatively new approach in the field of Machine Learning and has a large number of advantages to other conventional and popular classifiers like NNs. In this contribution we are using Least Square Support Vector Machines (LS-SVMs), which are the reformulations to the original SVMs.

A drawback of SVMs when dealing with audio data is their restriction to work with fixed-length vectors. Both in the kernel evaluation and in the simple input space dot product, the units under processing are vectors of constant size. However, when working with audio signals, although each signal frame is converted into a feature vector of a given size, the whole acoustic event is represented by a sequence of feature vectors, which shows variable length. In order to apply a SVM to this kind of data, one approach is to normalize the size of the sequence of input space feature vectors by extracting some statistical parameters (e.g. mean and standard deviation) from the sequence of vectors and thus transform the problem into that of fixed-length vector spaces. This is probably the most common and the easiest method. However, when frame-level features are transformed into statistical event-level features there is an unavoidable loss of information [14]. Another approach is to find a suitable kernel function that can deal with sequential data. In this work Fisher kernel, which make use of underlying generative model, is used to enable SVMs classify whole sequences.

The rest of this paper is organized as follows: next section explains the emotion database used in this research, then the structure of the AER system proposed in this work and the corresponding steps are demonstrated. The experimental results and the conclusion are presented after that.

THE EMOTION CORPUS

The database used in this research is the one created in [7].

This audio-visual emotion database is a professional reference database for testing and evaluating video, audio or joint audio-visual emotion recognition algorithms. The final version of the database contains 42 subjects, coming from 14 different nationalities. Among the 42 subjects, 81% are men, while the remaining 19% are women. First, each subject is asked to listen carefully to a short story for each of the six emotions (happiness, sadness, surprise, disgust, fear, and anger) and to immerge themselves into the situation. Once the subject is ready, he or she may read, memorize and pronounce the five proposed utterances (one at the time), which constitutes five different reactions to the given situation. The subjects are asked to put as much expressiveness as possible, producing a message that contains only the emotion to be elicited. All the subjects talk in English but they are from 14 different nationalities, so they might have different accents. All the utterances are approved by two experts in order to be genuine.

The sampling rate is 44100 Hz using two channels.

AER SYSTEM

The structure of the proposed speech emotion recognition system used in this paper is depicted in Fig. 1.

Preprocessing and windowing

In the preprocessing stage first each signal is de-noised by soft-thresholding the wavelet coefficients after three level of decomposition and since the silent parts of the signals do not carry any useful information, those parts including the leading and trailing edges are eliminated by thresholding the amount of energy in the small intervals of the signal. A Hamming window of length 23ms with 50% frequency overlap is used to divide signals into consequent frames.

Feature Extraction

Table 1 shows the list of features utilized in this work. While most researchers use prosodic feature and their statistical characteristics [2, 9, 5], we are proposing a set of novel features. Among these features only Mel Frequency Cepstrum Coefficients cation(MFCC) and Zero Crossing Rate (ZCR) have been used for speech emotion recognition in the past [3, 10, 8], while the rest are being used for the first time in this application.

While cepstral features and time-domain features are calculated from each frame, spectral features are extracted from non-overlapping logarithmically scaled frequency sub-bands listed in Table 2. The sub-band approach will provide better discrimination since for each emotion different energy distributions in different frequency sub-bands can be captured. Also because most of the important information in speech signals is located in the lower frequencies, the 6th sub-band is dismissed.

Classification

The recognition of human emotion is essentially a classification task. We are using Least Square-Support Vector Machines (LS-SVM) as a classifier in this research.

SVM is used in applications of regression and classification; however, it is mostly used as a binary classifier. SVM is based on the principle of structural risk minimization. The optimal boundary is found in such a way that maximizes the margin between two classes of data points [11]. SVM is based on kernel functions, which are used to map data points to a higher dimensional feature space in order to be linearly separable. In LS-SVM classifiers the original SVM formulation of Vapnik is modified by replacing inequality constraints with equality constraints. As a result the solution follows from solving a set of linear equations instead of a quadratic programming problem. The problem with SVMs is that they can not easily deal with the dynamic time structure of audio signals, since they are constrained to work with fixed-length vectors. There are several methods to cope with this problem [14] among which Fisher kernels has proven the most satisfying result. Fisher kernels make use of the information obtained by underlying generative models and enable us to classify the whole sequences rather than frame-level classification. The basic theory of Fisher kernels is to map the variable-length sequences to a single point in fixed-dimension space called score-space. The Fisher scores for a given input sequence $\mathbf{x}$ are computed as:

$$\mathbf{U_x} = \bigtriangledown_{\theta} log P(\mathbf{x}|\boldsymbol{\theta}) \qquad (1)$$

where $\boldsymbol{\theta}$ is the set of parameters of the generative model. When the model is considered Gaussian Mixture Model (GMM) $\boldsymbol{\theta}$ consists of mean vectors, covariance matrices, and weights.

Because we need to classify six different emotions, fuzzy-pairwise [12] method is used to extend the binary SVMs to a multi-category classifi problem.

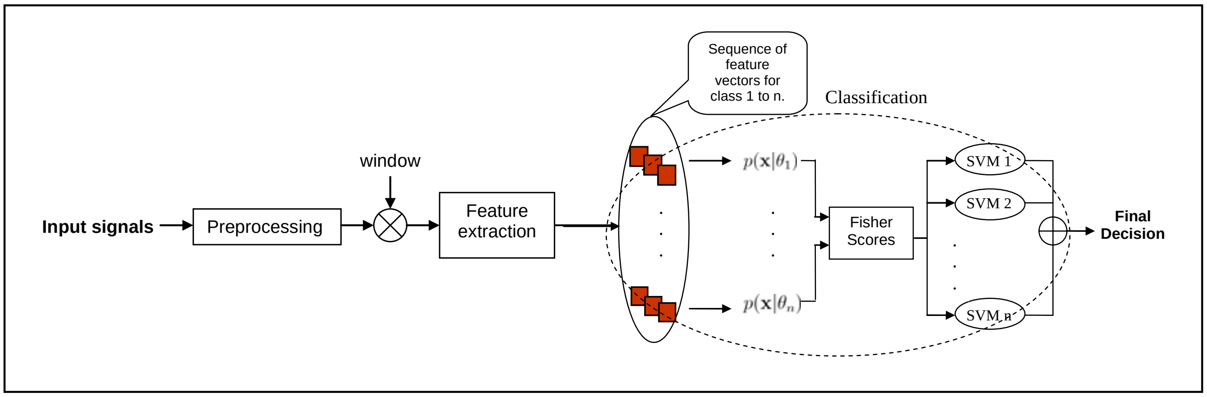

Figure 1. The structure of the speech emotion recognition

Table 1. List of acoustic features used for speech emotion recognition.

Audio Features		
Spectral Features	**Cepstral Features**	Time-domain Features
• Shannon Entropy • Renyi Entropy • Spectral Band Width • Spectral Centroid • Spectral Flux • Spectral Roll-Off Frequency • Spectral Flatness Measure • Spectral Crest Factor • Spectral Band Energy	MFCC	Zero Crossing Rate

Table 2. Sub-band allocation for calculating spectral features.

Sub-band	**Lower Edge (Hz)**	**Upper Edge (Hz)**
1	0	780
2	780	2000
3	2000	3900
4	3900	6800
5	6800	11500
6	11500	22050

IMPLEMENTED RESULTS

All the binary LS-SVMs are trained using linear kernel functions with different regularization parameters. 5-fold cross validation method is used to evaluate the performance of the trained classifiers. MATLAB LS-SVM toolbox [13] was used to implement LS-SVM classifiers. The overall accuracy of 97.65% were achieved which is a high accuracy comparing to other reported results in literature [9, 5, 2, 6]. The confusion matrix is presented in Table 3. As it shows the most difficult emotion to recognize in our experiment is surprise and also surprise and disgust have the highest probability to be confused with each other.

CONCLUSION

In this contribution, we introduced an Automatic Human Speech Emotion Recognition System using a set of novel acoustic features which most of them are used for the first time in the application of AER. Also instead of extracting the features from frames, they are calculated from different frequency sub-bands in order to capture more detail and therefore more discrimination power. For classification we used sequence discriminant SVMs using Fisher kernels which is proved to be a very successful method. To efficiently extend our binary classifiers to a multi-category problem, fuzzy-pairwise method was adopted. All these state-of –the-art approaches resulted in 97.65% accuracy which is a very satisfying and superior accuracy compare to previous works.

Table 3. Confusion matrix for recognized emotions.

	Recognized Emotions (%)					
	Ang	*Fea*	*Dis*	*Hap*	*Sad*	*Sur*
Ang	**99.85**	0	0.15	0	0	0
Fea	0	**100**	0	0	0	0
Dis	0	0	**99.85**	0	0.15	0
Hap	0.15	0	0	**99.85**	0	0
Sad	0	0	0.31	0	**99.69**	0
Sur	0	0	1.24	0	0	**98.76**

REFERENCES

[1] S. Hoch, F. Althoff, G. McGlaun, and G. Rigoll, “Bimodal fusion of emotional data in an automotive environment,” *Proceeding of IEEE International Conference on Acoustic, Speech, and Signal Processing,* Vol. 2, PP. 1085-1088, 18-23 March 2005.

[2] C.A. Martinez and A.B. Cruz, “Emotion recognition in non-structured utterance for human-robot interaction”, *IEEE International Workshop on Robot and Human Interactive Communication,* PP. 19-23, 13-15 Aug. 2005.

[3] T. Nguyen and I. Bass, “Investigation of combining SVM and decision tree for emotion classification*,” seventh IEEE International Symposium on Multimedia,* PP. 540-544, 2005.

[4] B. Schuller, G. Rigoll, and M. Lang, “Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture, ” *Proceedings of IEEE International Conference on Acoustic, Speech, and Signal Processing,* Vol.1, PP. I-577-80, 17-21 May, 2004.

[5] B. Schuller, G. Rigoll, and M. Lang, “Hidden Markov model-based speech emotion recognition, ” *Proceeding of the IEEE International Conference on Acoustic, Speech, and Signal Processing*, PP. I-401-04, 6-10 April, 2003.

[6] V. A. Petrushin, “Creating emotion recognition agents for speech signal, ” unpublished.

[7] O. Martin, I. Kotsia, B. Macq, and I. Pitas, “The eNTERFACE’05 audio-visual emotion database,” *Proceedings of the 22nd International Conference on Data Emgineering Workshop*, 3-7 April 2006.

[8] YL. Lin and G. Wei, “Speech emotion recognition based on HMM and SVM,” *Proceeding of International Conference on Machine Learning and Cybernetics,* Vol. 8, PP 4898-4901, 18-21 Aug. 2005.

[9] J. Nicholson, K. Takahashi, and R. Nakatsu, “Emotion recognition in speech using Nueral Networks,” *Proceedings of the 6th International Conference on Neural Information Processing*, vol. 2, PP. 495-501, 1999.

[10] ZJ. Chuang, CH. Wu, “Emotion recognition using acoustic features and textual content”, *IEEE International Conference on Multimedia and Expo,* Vol. 1, PP. 53-56, 27-3- June 2004.

[11] N. Cristianini and J. SH. Taylor, An Introduction to Support Vector Machines and Other Kernel-based Methods. United Kingdom: Cambridge University Press, 2000.

[12] D. Tsujinishi, Y. Koshiba, and SH. Abe, “Why pairwise is better that One-against-All or All-at-Once,” *Proceedings of IEEE International Conference on Neural Networks*, vol. 1, PP. 693-698, July 2004.

[13] K. Pelckmans, J. A.K. Suykens, T. V. Gestel, J. De Brabanter, L. Lukas, B. Hamers, B. De Moor, and J. Vandewalle, “LS-SVMlab toolbox”, Department of Electricla Engineering, Katholieke Universiteit Leuven, February 2003.

[14] A. Temko, E. Monte, and C. Nadeu, “Comparison of sequence discriminant support vector machines for acoustic event classification,” *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing,* Vol. 5, PP. v-721-v-724, May 2006.