Speech Emotion Recognition: Recurrent Neural Networks compared to SVM and Linear Regression

Leila Kerkeni, Youssef Serrestou, Mohamed Mbarki, Mohamed Mahjoub, Kosai Raoof, Catherine Cléder

Speech Emotion Recognition: Recurrent Neural Networks compared to SVM and Linear Regression

Leila Kerkeni[1,2], Youssef Serrestou[1], Mohamed Mbarki[2], Mohamed Ali Mahjoub[2], Kosai Raoof[1], and Catherine Cleder[3]

[1] Acoustics Laboratory of the University of Maine,
[2] Laboratory of Advanced Technology and Intelligent Systems,
[3] Research Centre for Education of the University of Nantes
kerkeni.leila@gmail.com

Abstract. Emotion recognition in spoken dialogues has been gaining increasing interest all through current years. A speech emotion recognition (SER) is a challenging research area in the field of Human Computer Interaction (HCI). It refers to the ability of detection the current emotional state of a human being from his or her voice. SER has potentially wide applications, such as the interface with robots, banking, call centers, car board systems, computer games etc. In our research we are interested to how, emotion recognition, can top enhance the quality of teaching for both of classroom orchestration and E-learnning. Integration of SER into aided teaching system, can guide teacher to decide what subjects can be taught and must be able to develop strategies for managing emotions within the learning environment. In linguistic activity, from student's interaction and articulation, we can extract information about their emotional state. That is why learner's emotional state should be considered in the language classroom. In general, the SER is a computational task consisting of two major parts: feature extraction and emotion machine classification. The questions that arise here: What are the acoustic features needed for a most robust automatic recognition of a speaker's emotion? Which methods is most appropriate for classification? How the database used influence the recognition of emotion in speech? Thus came the idea to compare a RNN method with the basic method (LR)[1] and the most widely used method (SVM). Most of previously published works generally use the berlin database. In this work we use another database. To our knowledge the spanish emotional database has never been used before. In recent years in speech emotion recognition, many researchers [2] proposed important speech features which contain emotion information and many classification algorithms. The aim of this paper is to compare firstly differents approachs that have proven their efficiency for emotions recognition task. Then to propose an efficient solution based on combination of these approachs. For classification, Linear regression (LR), Support vector machine (SVM) and Recurrent neural network (RNN) classifiers are used to classify seven different emotions present in the German and Spanish databases. The explored features included: mel-frequency cepstrum coefficients (MFCC)

[2] and modulation spectral features (MSFs) [3]. Table 1 show the recognition rate for each combination of various features and classifiers for Berlin and Spanish databases. The overall experimental results reveal that the feature combination of MFCC and MS has the highest accuracy rate on both Spanish emotional database using RNN classifier 90,05% and Berlin emotional database using LR 82,41%. These results can be explained as follows: LR classifier performed better results with feature combination of MFCC and MS for both databases. And under the conditions of limited training data (Berlin database), it can have a very good classification performance compared to other classifiers. A high dimension can maximize the rate of LR. As regarding the SVM method, we found the same results as these presented in [3]. The MS is the most appropriate features for SVM classifier. To improve the performance of SVM, we need to change the model for each types of features. To the spanish database, the feature combination of MFCC and MS using RNN has the best recognition rate 90.05%. For Berlin database, combination both types of features has the worst recognition rate. That because the RNN model having too many parameters (155 coefficients in total)and a poor training data. This is the phenomena of overfitting. The performance of SER system is influenced by many factors, especially the quality of samples, the features extracted and classification algorithms. Nowadays, a lot of uncertainties are still present for the best algorithm to classify emotions and what features influence the recognition of emotion in speech. To extract the more effective features of speech, seek for an efficient classification techniques and enhance the emotion recognition accuracy is our future work. More work is needed to improve the system so that it can be better used in classroom orchestration.

Table 1. Recognition results using RNN, SVM and LR classfiers based on Berlin and Spanish databases

Dataset	**Feature**	**RNN (%)**	**SVM (%)**	**LR (%)**
Berlin	MS	66.32	63.30	60.70
	MFCC	69.55	56.60	67.10
	MFCC+MS	58.51	59.50	75.90
Spanish	MS	82.30	77.63	70.60
	MFCC	86.56	70.69	76.08
	MFCC+MS	**90.05**	68.11	82.41

Keywords: Speech Emotion Recognition, Recurrent Neural Networks, SVM, Linear Regression, MFCC, Modulation Spectral Features.

References

1. Naseem, I., Togneri, R., Member, S., IEEE, Bennamoun., M.: Linear Regression for Face Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 32 (2010)
2. Surabhi, V., Saurabh, M.: Speech Emotion Recognition: A review. IRJET 03 (2016)
3. Wua, S., b, T.H.F., Chan, W.Y.: Automatic speech emotion recognition using modulation spectral features. Speech Communication 53: 768-785 (2011)