Speech Emotion Recognition: Methods and Cases Study

Conference Paper · January 2018

DOI: 10.5220/0006611601750182

5 authors, including:

Leila Kerkeni
Université du Maine
8 PUBLICATIONS **40** CITATIONS
SEE PROFILE

Youssef Serrestou
Université du Maine
24 PUBLICATIONS **93** CITATIONS
SEE PROFILE

Mohamed Mbarki
Universite de Sousse Tunisie
9 PUBLICATIONS **40** CITATIONS
SEE PROFILE

Kosai Raoof
Le Mans Université - France
137 PUBLICATIONS **729** CITATIONS
SEE PROFILE

Speech Emotion Recognition: Methods and Cases Study

Leila Kerkeni[1,2], Youssef Serrestou[1], Mohamed Mbarki[3], Kosai Raoof[1] and Mohamed Ali Mahjoub[2]
[1]*LAUM Acoustics Laboratory of the University of Maine, Le Mans University, France*
[2]*LATIS Laboratory of Advanced Technologies and Intelligent Systems, University of Sousse, Tunisia*
[3]*Higher Institute of Applied Sciences and Technology of Sousse, Univerisity of Sousse, Tunisia*

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

Abstract: In this paper we compare different approaches for emotions recognition task and we propose an efficient solution based on combination of these approaches. Recurrent neural network (RNN) classifier is used to classify seven emotions found in the Berlin and Spanish databases. Its performances are compared to Multivariate linear regression (MLR) and Support vector machine (SVM) classifiers. The explored features included: mel-frequency cepstrum coefficients (MFCC) and modulation spectral features (MSFs). Finally results for different combinations of the features and on different databases are compared and explained. 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 MLR 82,41%.

1 INTRODUCTION

Emotion recognition in spoken dialogues has been gaining increasing interest all through current years. Speech Emotion Recognition (SER) is a hot research topic in the field of Human Computer Interaction (HCI). It has a potentially wide applications, such as the interface with robots, banking, call centers, car board systems, computer games etc. For classroom orchestration or E-learning, information about the emotional state of students can provide focus on enhancement of teaching quality. For example teacher can use SER to decide what subjects can be taught and must be able to develop strategies for managing emotions within the learning environment. That is why learner's emotional state should be considered in the 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 is the optimal feature set? What combination of acoustic features for a most robust automatic recognition of a speaker's emotion? Which method is most appropriate for classification? Thus came the idea to compare a RNN method with the basic method MLR and the most widely used method SVM. And also all previously published works generally use the berlin database. To our knowledge the spanish emotional database has never been used before. For this reason we have chosen to compare them. In fact, the emotional feature extraction is a main issue in the SER system. Many researchers (Surabhi and Saurabh, 2016) have proposed important speech features which contain emotion information, such as energy, pitch, formant frequency, Linear Prediction Cepstrum Coefficients (LPCC), Mel-Frequency Cepstrum Coefficients (MFCC) and modulation spectral features (MSFs) (Wua et al., 2011). The last step of speech emotion recognition is classification. It involves classifying the raw data in the form of utterance or frame of the utterance into particular class of emotion on the basis of features extracted from the data. In recent years in speech emotion recognition, researchers proposed many classification algorithms, such as Gaussian Mixture Model (GMM)(Martin and Robert, 2009), Hidden Markov Model (HMM) (B. Ingale and Chaudhari, 2012), Support Vector Machine (SVM)(A. et al., 2013),(G.S. et al., 2016),(Pan et al., 2012), (Peipei et al., 2011), Neural Networks (NN) (Sathit, 2015) and Recurrent Neural Networks (RNN) (Alex and Navdeep, 2014), (Lim et al., 2017), (Chen and Jin, 2015). Some other types of classifiers are also proposed by some researchers such as a modified brain emotional learning model (BEL) (Sara et al., 2017) in which the Adap-

Kerkeni, L., Serrestou, Y., Mbarki, M., Raoof, K. and Mahjoub, M.
Speech Emotion Recognition: Methods and Cases Study.
DOI: 10.5220/0006611601750182
In *Proceedings of the 10th International Conference on Agents and Artificial Intelligence (ICAART 2018) - Volume 2*, pages 175-182
ISBN: 978-989-758-275-2

tative Neuro-Fuzzy Inference System (ANFIS) and Multilayer Perceptron (MLP) are merged for speech emotion recognition. Another proposed strategy is a multiple kernel Gaussian process (GP) classification (Chen and Jin, 2015), in which two similars notions in the learning algorithm are presented by combining the linear kernel and radial basis function (RBF) kernel. The Voiced Segment Selection (VSS) algorithm also proposed in (Yu et al., 2016) deals with the voiced signal segment as the texture image processing feature which is different from the traditional method. It uses the Log-Gabor filters to extract the voiced and unvoiced features from spectrogram to make the classification. Speech emotion recognition is essentially a sequence classification problem, where the input is a variable-length sequence and the output is one single label. That is why we have chosen recurrent neural networks in our work. In this experimental work, we have used Multivariate Linear Resgression (MLR), Support Vector Machine (SVM) and Recurrent Neural Networks (RNN) classifiers to identify the emotional state of spoken utterances. In order to demonstrate the high effectivennes of the MFCC and MS features extraction for emotion classification in speech, we provide results on two open emotional databases (Berlin-DB and Spanish-DB).

The remainder of the paper is organized as follows: Section 2 describes the databases used in the experiments. The speech features as presented in section 3. The several classification methods used in our work are introduced in section 4. Experiments and results are performed in section 5, and conclusion follows in section 6.

2 EMOTIONAL SPEECH DATA

The performance and robustness of the recognition systems will be easily affected if it is not well-trained with suitable database. Therefore, it is essential to have sufficient and suitable phrases in the database to train the emotion recognition system and subsequently evaluate its performance. In this section, we detail the two emotional speech databases used in our experiments: Berlin Database and Spanish Database.

2.1 Berlin Emotional Speech Database

The Berlin database (Burkhardt et al., 2005) is widely used in emotional speech recognition. It contains 535 utterances spoken by 10 actors (5 female, 5 male) in 7 simulated emotions (anger, boredom, disgust, fear, joy, sadness and neutral). This Dataset was chosen for the following reasons: i) the quality of its recording is very good and ii) it is public (Ber,) and popular Dataset of emotion recognition that is recommended in the literature (Sara et al., 2017).

2.2 Spanish Emotional Database

The INTER1SP Spanish emotional database contains utterances from two profesional actors (one female and one male speackers).The spanish corpus that we have the right to access (free for academic and research use) (Spa,), was recorded twice in the 6 basic emotions plus neutral (anger, sadness, joy, fear, disgust, surprise, Neutral/normal). Four additional neutral variations (soft, loud, slow and fast) were recorded once. This is preferred to other created database because it is available for researchers use and it contains more data (4528 utterances in total). This paper has focused on only 7 main emotions from the Spanish Dataset in order to achieve a higher and more accurate rate of recognition and to make the comparison with the Berlin database detailed above.

3 FEATURE EXTRACTION

The speech signal contains a large number of parameters that reflect the emotional characteristics. One of the sticking points in emotion recognition is what features should be used. In recent research, many common features are extracted, such as energy, pitch, formant, and some spectrum features such as Linear Prediction Coefficients (LPC), Mel-Frequency Cepstrum Coefficients (MFCC) and Modulation spectral features. In this work, we have selected Modulation spectral features and MFCC, to exract the emotional features.

3.1 MFCC Features

Mel-Frequency Cepstrum coefficient is the most used representation of spectral property of voice signals. These are the best for speech recognition as it takes human perception sensitivity with respect to frequencies into consideration. For each frame, the Fourier transform and the energy spectrum were estimated and mapped into the mel-frequency scale. The Discrete Cosine Transform (DCT) of the mel log energies were estimated and the first 12 DCT coefficients provided the MFCC values used in the classification process. Usually, the process of calculating MFCC is shown in Figure 1.

In our research, we extract the first 12-order of the MFCC coefficients where the speech signals are sampled at 16 KHz. For each order coefficients, we

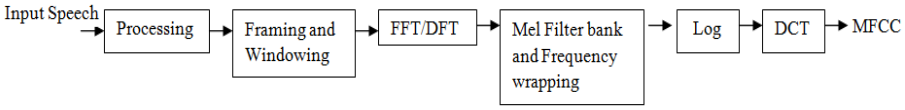

Figure 1: Schema of MFCC extraction (Srinivasan et al., 2014).

compute the mean, standard deviation, Kurtosis and Skewness, and this is for the other all the frames of an utterance. Each MFCC feature vector is 60-dimensional.

3.2 Modulation Spectral Features

Modulation spectral features (MSFs) are extracted from an auditory-inspired long-term spectro-temporal representation. These features are obtained by emulating the Spectro-temporal (ST) processing performed in the human auditory system and considers regular acoustic frequency jointly with modulation frequency. The steps for computing the ST representation are illustrated in figure 2.In order to obtain the ST representation, the speech signal is first decomposed by an auditory filterbank. The Hilbert envelopes of the critical-band outputs are computed to form the modulation signals. A modulation filterbank is further applied to the Hilbert envelopes to perform frequency analysis. The spectral contents of the modulation signals are referred to as modulation spectra, and the proposed features are thereby named modulation spectral features (MSFs) (Wua et al., 2011). Lastly, the ST representation is formed by measuring the energy of the decomposed envelope signals, as a function of regular acoustic frequency and modulation frequency. The mean of energy, taken over all frames in every spectral band provides a feature. In total, 95 MSFs are calculated in this work from the ST representation.

4 CLASSIFICATION

4.1 Multivariate Linear Regression Classification

Multivariate Linear Regression (MLR) is a simple and efficient computation of machine learning algorithms, and it can be used for both regression and classification problems. We have slightly modified the LRC agorithm described as follow 1 (Naseem et al., 2010). We calculated (in step 3) the absolute value of the difference between original and predicted response vectors ($\mid y - y_i \mid$), instead of the euclidean distance between them ($\mid\mid y - y_i \mid\mid$):

Algorithm 1 : Linear Regression Classification (LRC)

Inputs: Class models $X_i \in \mathbb{R}^{q \times p_i}$, $i = 1,2,...,N$ and a test speech vector $y \in \mathbb{R}^{q \times 1}$
Output: Class of y

1. $\hat{\beta}_i \in \mathbb{R}^{p_i \times 1}$ is evaluated against each class model,$\hat{\beta}_i = (X_i^T X_i)^{(-1)} X_i^T y$, $i = 1,2,...,N$
2. $\hat{y}_i$ is computed for each $\hat{\beta}_i, \hat{y}_i = X_i \hat{\beta}_i, \quad i = 1,2,...,N$;
3. Distance calculation between original and predicted reponse variables $d_i(y) = \mid y - y_i \mid, \quad i = 1,2,...,N$;
4. Decision is made in favor of the class with the minimum distance $d_i(y)$

4.2 Support Vector Machine

Support Vector Machines (SVM) is an optimal margin classifiers in machine learning. It is also used extensively in many studies that related to audio emotion recognition which can be found in (A. et al., 2013), (Peipei et al., 2011) and (Pan et al., 2012). It can have a very good classification performance compared to other classifiers especially for limited training data (G.S. et al., 2016). SVM theoretical background can be found in (Gunn, 1998). A MATLAB toolbox implementing SVM is freely available in (Too,).

4.3 Recurrent Neural Networks

Recurrent Neural Networks (RNN) are suitable for learning time series data. While RNN models are effective at learning temporal correlations, they suffer from the vanishing gradient problem which increases with the length of the training sequences. To resolve this problem, LSTM (Long Short Term Memory) RNNs were proposed by Hochreiter et al (Sepp and Jurgen, 1997) it uses memory cells to store information so that it can exploit long range dependencies in the data (Chen and Jin, 2015).

Figure 3 shows a basic concept of RNN implementation. Unlike traditional neural network that uses different parameters at each layer, the RNN shares the

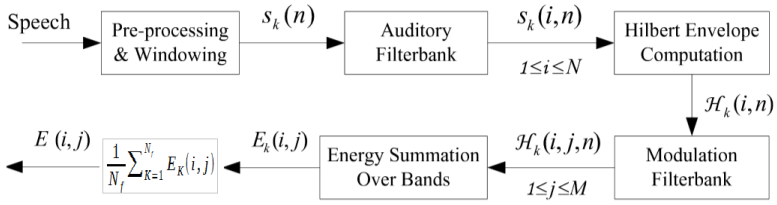

Figure 2: Process for computing the ST representation (Wua et al., 2011).

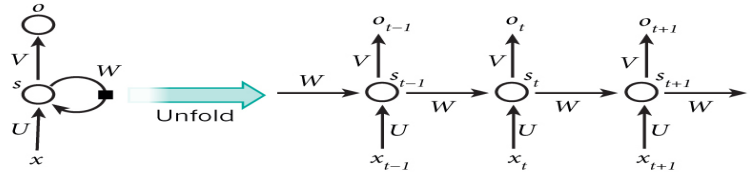

Figure 3: A basic concept of RNN and unfolding in time of the computation involved in its forward computation (Lim et al., 2017).

same parameters (U,V and W in figure 3) across all steps. The hidden state formulas and variables are as follows:

$$s_t = f(Ux_t + Ws_{t-1}) \quad (1)$$

with:

- x_t, s_t and o_t are respectively the input, the hidden state and the output at time step t;
- U, V, W are parameters matrices.

5 EXPERIMENTAL RESULTS

In this section, we describe the experiment environment and report the recognition accuracy of using MLR, SVM and RNN classifiers on two emotional speech database. We used Berlin database and spanish database for network training and validation. To evaluate the classification error 10-cross validation test were used. We used 70% of data for training and 30 % for testing. The neural network structure used is a simple LSTM. It consists of two consecutive LSTM layers with hyperbolic tangent activations followed by two classification dense layers. More detailed diagrams are shown in figure 4, 5 and 6 and can be found in appendix A. Table 1, 2 and 3 show the recognition rate for each combination of various features and classifiers based on Berlin and spanish databases.

As shown in table 1, MLR 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 MLR.

As regarding the SVM method, we found the same results as these presented in (Wua et al., 2011). The MS features achieve the best accuracy using 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 confusion matrix for recognition of emotions using MFCC and MS features with RNN based on spanish database is show in Table 4. The rate column lists per class recognition rates, and precision for a class is the number of samples correctly classified divided by the total number of samples classified to the class. It can be seen that *Sadness* was the emotion that was least difficult to recognize from speech as opposed to *Neutral* which was the most difficult and it forms the most notable confusion pair with *sadness*.

6 CONCLUSION AND FUTURE WORK

A lot of uncertainties are still present for the best algorithm to classify emotions. Different combinations of emotional features give different emotion detection rate. The researchers are still debating for what features influence the recognition of emotion in speech. In this article, the best result of recognition rate was **90.05 %**, achieved by combining the MFCC and MS

Table 1: Recognition results using MLR classifier based on Berlin and Spanish databases.

database	Features		A	E	F	L	N	T	W	Rate (%)
Berlin	MS	avg	41,79	29,86	42,92	75,40	54,84	85,64	78,10	**60,70**
		σ	10,97	9,86	9,07	10,85	6,63	13,37	8,40	2,50
	MFCC	avg	54,48	61,77	46,56	52,05	64,61	80,54	92,67	**67,10**
		σ	19,22	16,82	9,07	10,69	8,47	14,72	7,17	3,96
	MFCC+MS	avg	83,63	67,18	56,05	79,43	75,20	87,59	78,92	**75,90**
		σ	9,40	26,43	15,63	14,65	7,55	11,39	7,50	3,63
			A	D	F	J	N	S	T	Rate (%)
Spanish	MS	avg	61,61	53,08	72,42	54,20	90,97	61,59	68,16	**70,60**
		σ	3,70	4,03	4,29	4,67	2,14	3,90	4,62	1,37
	MFCC	avg	70,33	52,59	79,18	48,16	96,47	78,00	73,70	**76,08**
		σ	5,22	6,27	2,45	4,51	0,78	4,24	3,53	1,44
	MFCC+MS	avg	77,46	76,31	83,39	66,56	97,14	80,96	84,99	**82,41**
		σ	3,26	2,93	2,47	3,68	1,19	4,81	4,95	4,14

Spanish (a:anger, d:disgust, f:fear, j:joy, n:neutral, s:surprise, t: sadness) Berlin (a:fear, e:disgust, f:happiness, l:boredom, n:neutral, t:sadness, w:anger).

Table 2: Recognition results using SVM classifier based on Berlin and Spanish databases.

database	Features		A	E	F	L	N	T	W	Rate (%)
Berlin	MS	avg	60,35	57,54	49,75	66,54	62,93	80,02	67,01	**63,30**
		σ	12,55	22,72	18,14	13,90	12,70	9,36	8,40	4,99
	MFCC	avg	62,76	51,37	44,72	39,25	49,40	66,26	72,20	**56,60**
		σ	16,78	9,03	10,15	14,58	15,12	15,59	7,97	4,88
	MFCC+MS	avg	55,04	49,82	44,61	71,60	55,68	70,11	65,42	**59,50**
		σ	12,81	22,16	14,56	15,58	16,30	12,57	10,01	5,76
			A	D	F	J	N	S	T	Rate (%)
Spanish	MS	avg	71,99	68,72	79,54	65,59	86,93	69,76	79,76	**77,63**
		σ	6,45	4,21	3,15	5,86	3,50	3,60	3,78	1,67
	MFCC	avg	81,54	80,67	80,18	68,92	68,69	67,12	86,65	**70,69**
		σ	5,56	4,92	8,61	18,57	22,18	29,23	4,07	12,66
	MFCC+MS	avg	76,41	85,39	69,76	76,03	53,31	64,40	84,59	**68,11**
		σ	6,65	3,80	3,10	2,50	23,70	2,25	3,27	11,55

Spanish (a:anger, d:disgust, f:fear, j:joy, n:neutral, s:surprise, t: sadness) Berlin (a:fear, e:disgust, f:happiness, l:boredom, n:neutral, t:sadness, w:anger).

Table 3: Recognition results using RNN classifier based on Berlin and Spanish databases.

Dataset	**Feature**	**Average (avg)**	**Standard deviation (σ)**
Berlin	MS	66.32	5.93
	MFCC	69.55	3.91
	MFCC+MS	58.51	3.14
Spanish	MS	82.30	2.88
	MFCC	86.56	2.80
	MFCC+MS	**90.05**	1.64

Table 4: Confusion matrix for using MFCC and MS features based on spanish database.

Emotion	Anger	Disgust	Fear	Joy	Neutral	Surprise	Sadness	Rate (%)
Anger	**131**	14	3	23	8	2	0	72,38
Disgust	3	**197**	1	6	6	6	2	89,95
Fear	3	15	**115**	6	12	0	0	76,16
Joy	8	4	1	**411**	0	11	0	89,14
Neutral	9	14	9	4	**144**	1	1	79,12
surprise	1	4	0	18	0	**133**	0	85,26
Sadness	8	1	18	11	17	0	**93**	62,84
Precision (%)	80,37	79,12	78,23	85,80	77,00	86,92	96,87	

features and for the RNN model in the Spanish emotional database. Moreover, higher accuracy can be obtained using the combination of more features. Apart from this, seeking for robust feature representation is also considered as part of the ongoing research, as well as efficient classification techniques for automatic speech emotion recognition.

Methods based on the Fourier transform such as MFCC and MS are the most used in speech emotion recognition. However, their popularity and effectiveness have a downside. It has led to a very specific and limited view of frequency in the context of signal processing. Simply put, frequencies, in the context of Fourier methods, are just a collection of the individual frequencies of periodic signals that a given signal is composed of. To use methods that it provides an alternative interpretation of frequency and an alternative view of non-linear and non-stationary phenomena is our future work. More work is needed to improve the system so that it can be better used in real-time speech emotion recognition.

REFERENCES

Berlin database of emotional speech. http://emodb.bilderbar.info/start.html.

Berlin database of emotional speech. http://www.elra.info/en/catalogues/catalogue-language-resources/.

Svm and kernel methods matlab toolbox. http://asi.insa-rouen.fr/enseignants/ arakoto/toolbox/.

A., M., S., S. R., and S., T. S. (2013). SVM Scheme for Speech Emotion Recognition Using MFCC Feature. *International Journal of Computer Applications*, 69.

Alex, G. and Navdeep, J. (2014). Towards end-to-End Speech Recognition with Recurrent Neural Networks. *International Conference on Machine Learning*, 32.

B. Ingale, A. and Chaudhari, D. (2012). Speech Emotion Recognition Using Hidden Markov Model and Support Vector Machine. *International Journal of Advanced Engineering Research and Studies*.

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W., and Weiss, B. (2005). A Database of German Emotional Speech. *INTERSPEECH*.

Chen, S. and Jin, Q. (2015). Multi-modal Dimensional Emotion Recognition using Recurrent Neural Networks. Brisbane, Australia.

G.S., D. S., P., C., and B., V. (2016). SVM Based Speech Emotion Recognition Compared with GMM-UBM and NN. *IJESC*, 6.

Gunn, S. R. (1998). *Support Vector Machines for Classification and Regression*. PhD thesis.

Lim, W., Jang, D., and Lee, T. (2017). Speech Emotion Recognition using Convolutional and Recurrent Neural Networks. Asia-Pacific.

Martin, V. and Robert, V. (2009). Recognition of Emotions in German Speech Using Gaussian Mixture Models. *LNAI 5398*, pages 256–263.

Naseem, I., Togneri, R., Member, S., IEEE, and Bennamoun., M. (2010). Linear Regression for Face Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32.

Pan, Y., Shen, P., and Shen, L. (2012). Speech Emotion Recognition Using Support Vector Machine. *International Journal of Smart Home*, 6.

Peipei, S., zhou, C., and Xiong, C. (2011). Automatic speech Emotion Recognition Using Support Vector Machine.

Sara, M., Saeed, S., and Rabiee, A. (2017). Speech emotion Recognition Based on a Modified Brain Emotional Learning Model. *Elsevier*, pages 32–38.

Sathit, P. (2015). Improvement Of Speech Emotion Recognition with Neural Network Classifier by Using Speech Spectrogram.

Sepp, H. and Jurgen, S. (1997). Long Short-term Memory. *Neural Computation*.

Surabhi, V. and Saurabh, M. (2016). Speech Emotion Recognition: A review. *IRJET*, 03.

V. Srinivasan, V. Ramalingam, and P. Arulmozhi (2014). Artificial Neural Network Based Pathological Voice Classification Using Mfcc Features.

Wua, S., b, T. H. F., and Chan, W.-Y. (2011). Automatic speech emotion recognition using modulation spectral features. *Speech Communication 53: 768-785*.

Yu, G., Eric, P., Hai-Xiang, L., and van den, H. J. (2016). Speech emotion Recognition Using Voiced Segment Selection Algorithm.

APPENDIX: LSTM NETWORK

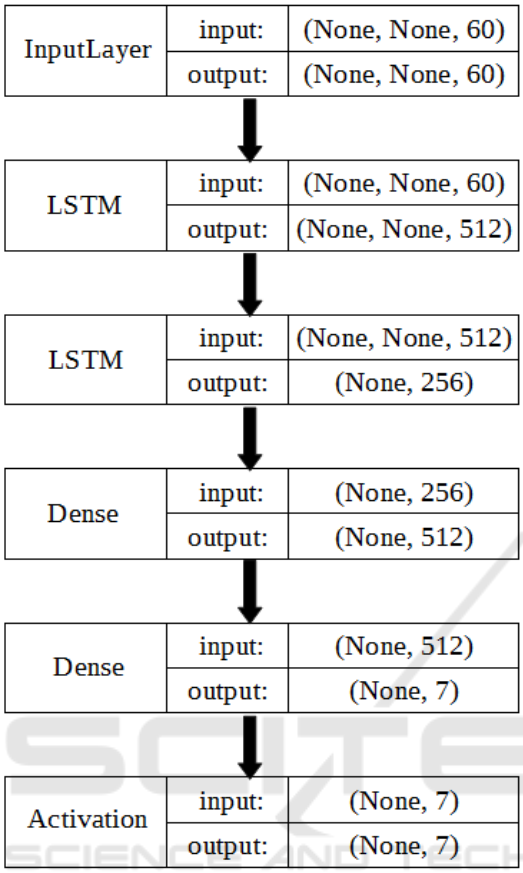

Figure 4: LSTM network architecture using MFCC features.

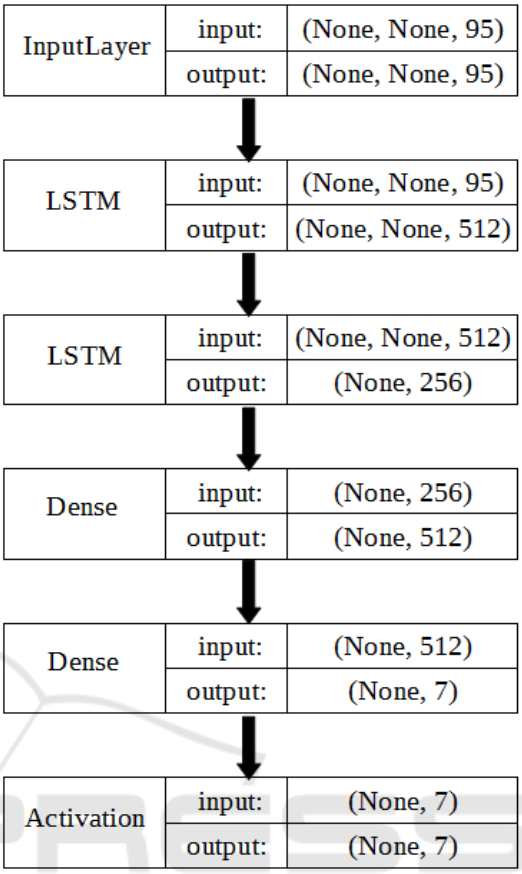

Figure 5: LSTM network architecture using MS features.

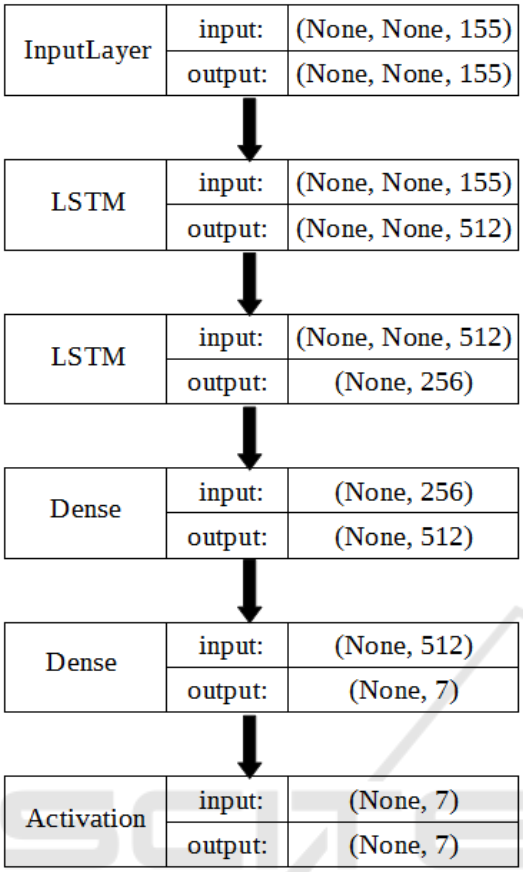

Figure 6: LSTM network architecture using combination of MFCC and MS features.