Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 00 (2014) 000–000

Procedia Computer Science

www.elsevier.com/locate/procedia

6th International conference on Intelligent Human Computer Interaction, IHCI 2014

Speech emotion recognition in acted and spontaneous context

Farah Chenchah[a]*, Zied Lachiri [a]

[a]*LR-SITI Laboratory*
National Institute of Applied Science and Technology,BP.676 centre urbain cedex Tunis, Tunisia

Abstract

Little attention has been paid so far in the context in which databases used for the study of emotion through vocal channel are recorded. Thus, we propose and evaluate an emotion classification system focusing on the differences between acted and spontaneous emotional speech through the use of two different databases: SAVEE and IEMOCAP. For the purpose of this work, we have examined wavelet packet energy and entropy features applied to Mel, Bark and ERB scale applied with Hidden Markov Model (HMM) as classification system. Experimental results show that the proposed method is a feasible technique for emotion classification for both acted and spontaneous context, pointing out the performance difference of the system between the two contexts. The experimental results shows that ERB scale features gives better performance in comparison with other studied features with recognition accuracy of 78.75% for acted context and 50.06% for spontaneous context.

Peer-review under responsibility of the Scientific Committee of IHCI 2014.

Keywords: Hidden Markov Model (HMM); emotion recognition; wavelet packet transforms; acted context; spontaneous context

1. Introduction

Emotions have a great impact on human behavior, since they influence processes such as perception, attention, learning, memory or decision-making. Speech signals convey not only words and meanings but also emotions [1]. A challenging research issue that has been of growing importance in the two last decades is to detect human emotion from several channels and particularly from voice.

* *E-mail address:* farahchenchah@yahoo.fr

Peer-review under responsibility of the Scientific Committee of IHCI 2014.

A number of recent studies have investigated the link between vocal channel and emotion detection, and a wide variety of databases were registered using several languages and different situations [2, 3].

The machine learning algorithms used in emotion recognition systems are trained and tested with data that describes the problem at hand. Therefore, the quality of emotional databases is extremely important [4]. Automatic emotion recognition from spontaneous speech is challenging due to non-ideal recording conditions and highly ambiguous ground truth labels. The main advantage of recording acted databases is that many aspects of the recording can be carefully and systematically controlled [5].

Several speech emotion recognition researches focused on acted context[6], others treated spontaneous context[7], but few of them have focused on studying the impact of the context on the system accuracy.

The use of a set of features to increase the recognition rate for speech emotion recognition have been widely adopted by researchers[8], in this field Schuller et al.[9] have used pitch and energy features, Kwon et al.[10] used pitch, log energy, formant and MFCC features and Lee et al.[11]combined pitch, energy, duration and formant. In the automatic speech recognition, wavelet features are frequently used and gives a promising results [12].

In this work, we implemented a speech emotion recognition system in which we used wavelet packet energy and entropy features applied to Mel, Bark and ERB scale. Classification was carried out through Hidden Markov Model (HMM) in order to use its characteristics of capturing the temporal activity incorporated in speech. We have particularly focused on the data quality issue and provided a comparison of emotion detection through vocal channels with acted and spontaneous context.

This paper is organized as follows: section 2 gives details of emotion recognition system architecture where an overview of wavelet packet decomposition is detailed followed by Hidden Markov Models introduced as classifiers. Section 3 shows experimental set up which gives details of the two emotional databases and output results of the work performed. Finally, conclusions of the experiment are drawn and we suggest possible future work.

2. System description

The framework of our approach is illustrated in figure 1. In the step of feature extraction, we use energy and entropy features applied to wavelet filter bank coefficients using Mel scale, Bark scale and ERB scale. The data were divided into two sets: training and testing. The training set was treated using the Hidden Markov Model as classifier. The accuracy of the system was then set based on the testing set.

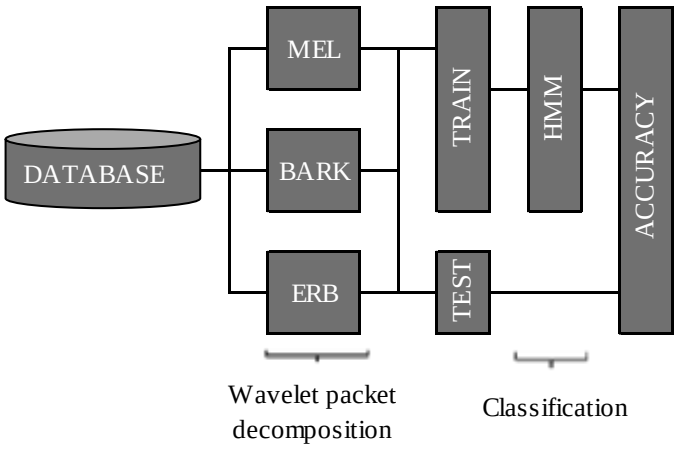

Fig 1.Speech emotion recognition system description

2.1. Feature extraction using wavelet packet decomposition

Wavelet Transform (WT) provides a linear powerful representation of signals. It gives time-frequency representation of the non-stationary signal. It decomposes signal over scaled and translated wavelets [13].

Discrete Wavelet Transform (DWT) is a multiresolution representation of a signal which decomposes signals into basis functions. A signal is decomposed into two frequency bands such as lower frequency band (approximation coefficients) and higher frequency band (detail coefficients). DWT is a left recursive binary tree structure which filtering process is iterated only for the approximation sub-band at each level of decomposition [14].

Wavelet Packets (WPs) are considered to have important signal representation schemes impacting compression, detection and classification. Wavelet Packet Transform (WPT) can be implemented through iterative decomposition of all coefficients, yielding an equal frequency bandwidth. In Wavelet Packet (WP) decomposition procedure, both lower and higher frequency bands are decomposed into two sub-bands. Each subspace is indexed by its depth i and the number of subspaces p:

$$\psi_{i+1}^{2p}(k) = \sum_{n=-\infty}^{+\infty} g[n]\psi_i^p(k-2^i n) \tag{1}$$

$$\psi_{i+1}^{2p+1}(k) = \sum_{n=-\infty} h[n]\psi_i^p(k-2^i n) \tag{2}$$

Where g[n] is a low pass filter and h[n] is the high pass filter.

Wavelet Packets can be used to characterize a rich covering of signal-space decomposition, and in particular, they provide a way for generating sub band dependent partitions of the observation space. In conclusion, WPs induce a family of structural filter-banks with a rich covering of time–frequency characteristics.

Wavelet packet (WP) based features are proposed, in which the decomposition closely follows the Mel scale; Bark scale and ERB scale [15,16,17]. The center frequency obtained of each filter using wavelet packet decomposition is given in table 1. The speech signals sampled at 16 kHz are filtered with the 24 Mel scale wavelet packet filters, 21 Bark scale wavelet packet filters and 24 ERB scale wavelet packet filters.

After performing the decomposition by WP of a signal, energy and entropy in each of the frequency bands have been calculated by:

$$Energy_j = \frac{\sum_{i=1}^{N_j}\left|W_j^P x(i)\right|^2}{N_j} \tag{3}$$

$$Entropy_j = -\sum_{i=1}^{N_j}\left|W_j^P\, x(i)\right|^2 log\left|W_j^P\, x(i)\right|^2 \tag{4}$$

j=1,2,.....,L.

Where W_j^p x(i) is the i-the coefficient of the wavelet packet transform of a signal x at node W_j^p of the wavelet packet. L is total number of nodes used (total sub-bands), N_j is total number of coefficients t node j (j-th sub-band).

Table 1 Frequency bands obtained from wavelet packet decomposition

MEL SCALE					
Critical band	Mel scale	Wavelet subband	Critical band	Mel scale	Wavelet subband
1	100	125	13	1 516	1 750
2	200	250	14	1 741	2 000
3	300	375	15	2 000	2 250
4	400	500	16	2 297	2 500
5	500	625	17	2 639	2 750
6	600	750	18	3 031	3 000
7	700	875	19	3 482	3 500
8	800	1 000	20	4 000	4 000
9	900	1 125	21	4 595	5 000
10	1 000	1 250	22	5 278	6 000
11	1 149	1 375	23	6 063	7 000
12	1 320	1 500	24	6 954	8 000

BARK SCALE					
Critical band	Bark scale	Wavelet subband	Critical band	Bark scale	Wavelet subband
1	50	125	13	1 850	2 250
2	150	250	14	2 150	2 500
3	250	375	15	2 500	3 000
4	350	500	16	2 900	3 500
5	450	625	17	3 400	4 000
6	570	750	18	4 000	5 000
7	700	875	19	4 800	6 000
8	840	1 000	20	5 800	7 000
9	1 000	1 250	21	7 000	8 000
10	1 170	1 500			
11	1 370	1 750			
12	1 600	2 000			

ERB SCALE					
Critical band	ERB scale	Wavelet subband	Critical band	ERB scale	Wavelet subband
1	50	62,5	13	1 285,9	1 250
2	92,2	125	14	1 515,4	1 500
3	140,9	187,5	15	1 779,5	1 750
4	196,9	250	16	2 083,7	2 000
5	261,3	312,5	17	2 434,0	2 500
6	335,6	375	18	2 837,3	3 000
7	421,1	437,5	19	3 301,7	3 500
8	519,5	500	20	3 836,4	4 000
9	632,8	625	21	4 452,2	5 000
10	763,4	750	22	5 161,2	6 000
11	913,6	875	23	5 977,6	7 000
12	1 086,7	1 000	24	6 917,6	8 000

2.2. HMM classification

Classification models to be settled aim to deliver useful information for segmenting the signals in some meaningful units. Several statistical classifiers dealing with high dimensional data have been used for emotion recognition models like support vector machines, neural networks, decision trees, and Hidden Markov model (HMM). However, HMM are most commonly found in the literature on emotional speech recognition [18]. Hidden Markov model (HMM) and their various forms (discrete, continuous, and semi-continuous) have been applied to speech recognition problems in general and speech emotion recognition in particular.

HMM is a Markov process that is split into two components: an observable component and an unobservable or 'hidden' component. It consists of five components: number of hidden states, number of observation symbols per state, state transition probability distribution, observation symbol probability distribution in each state and initial state probability distribution.

The training phase aims to determine model parameters of the HMMs based on the training data. These parameters include means and covariance of the state output probability distributions and probabilities of the state transitions.

3. Experimental Setup and results

3.1. Emotionnal databases

3.1.1. SAVEE

Surrey Audio-Visual Expressed Emotion (SAVEE) database consists of recordings from four male actors (DC, JE, JK, DC) in seven different emotions (fear, anger, disgust, sadness, surprise, happiness, neutral) [19]. Recordings consisted of 15 phonetically-balanced TIMIT sentences per emotion (with additional 30 sentences for neutral state) resulting into corpus of 480 British English utterances. The data were recorded by painting 60 markers on the face of actor for extraction of visual features.

3.1.2. IEMOCAP

IEMOCAP is an Interactive Emotional database collected at SAIL lab at USC [20]. It contains approximately twelve hours of audio-visual data from five mixed gender pairs of actors .There are five sessions, each recorded session lasts approximately five minutes and consists of two actors interacting with each other's. Two acting styles were used: improvisation of scripts and improvisation of hypothetical scenarios. The dyadic sessions were manually segmented into utterances .The emotional content of each utterance was annotated by human annotators in categorical labels :{angry, happy, sad, neutral, frustrated, excited, fearful, surprised, disgusted, other} and in terms of dimensional descriptions over the axes of :{valence, activation, dominance}.

3.2. Experimental setup

In our studies, we collect all the available sentences which are classified in four emotional states; angry, happy, neutral and sad. SAVEE corpus contains 240 utterances and IEMOCAP corpus contains 2218 sentences. Data were recorded at a sampling rate of 16KHz. The signal samples are segmented into frames of 50 ms each with 25 ms overlap between consecutive frames. For performing frequency partitioning using wavelet packet decomposition, we used 8th order of the Daubechies wavelet.

A feature database is created, after the computation of energy and entropy measures from each sub band wavelet packet coefficients and they are used as input features for the classifiers to distinguish the emotion state.

We first evaluate the topology of the HMM based classifiers by varying the number of states and the number of mixture components per state for various feature sets. We select the Bakis model with five states from which three are hidden with one Gaussian mixture. HMM models are built for four emotions individually.

In each database, 70% of the samples were used as training set, and 30% of the samples were used as test set.

The training and testing of HMM classifiers were performed using the Hidden Markov Toolkit (HTK) [21].

3.3. Results

Speech emotion recognition is implemented using wavelet packet energy and entropy features following Mel, Bark and ERB scale, emotion is recognized using HMM classifiers. We evaluate the system with two corpuses which represents two different contexts, acted and spontaneous.

Table 2 shows the classification results obtained from energy and entropy features applied to wavelet filter bank coefficient using MEL scale, Bark scale and ERB scale. The average accuracy over the sets of four speakers (DC, JE, JK, KL) applied with energy features shows that ERB scale gives the better results with a classification result of 78.75%, followed by MEL scale with 70% and BARK scale with 68.75%.The same experiment conducted using entropy features conduct to the same conclusion: entropy feature applied to ERB scale gives the best accuracy with 78.75%. Furthermore, ERB applied with energy and entropy features still give the best results for three of the four speakers.

Table 2.Classification accuracy for energy and entropy features (SAVEE)

	MEL		**BARK**		**ERB**	
	Energy	Entropy	Energy	Entropy	Energy	Entropy
DC	75,00%	70,00%	75,00%	65,00%	85,00%	85,00%
JE	75,00%	80,00%	75,00%	85,00%	90,00%	95,00%
JK	60,00%	60,00%	70,00%	70,00%	75,00%	75,00%
KL	70,00%	65,00%	55,00%	50,00%	65,00%	60,00%
MEAN	**70,00%**	**68,75%**	**68,75%**	**67,50%**	**78,75%**	**78,75%**

Table 3 shows the accuracy obtained from classification results obtained by applying HMM classifier to IEMOCAP database. The classification accuracy obtained using energy features carry in the range of [34.85%, 61.02%] depending on the speaker and on the feature scale used. The best average accuracy from the data set is obtained with wavelet filter bank using ERB scale, with a mean of classification rate of 48.87%. As presented for SAVEE data base, entropy features were applied to the set of ten speakers of the IEMOCAP. The entropy features accuracy gives the best result when wavelet filter banks follow ERB scale with 50.06%. Moreover, we observe that entropy features in this scale outperform energy features for the same scale. However, energy features results using Mel scale and Bark scale exceeds those obtained with entropy features.

Table 3.Classification accuracy for energy and entropy features (IEMOCAP)

	MEL		**BARK**		**ERB**	
	Energy	Entropy	Energy	Entropy	Energy	Entropy
SP1	45,45%	38,96%	42,86%	38,96%	42,86%	55,84%
SP2	39,66%	39,66%	38,79%	39,66%	41,38%	49,14%
SP3	38,81%	34,33%	43,28%	40,30%	44,78%	41,79%
SP4	53,13%	52,08%	48,96%	58,33%	59,38%	65,63%
SP5	45,54%	44,55%	42,57%	48,51%	46,53%	47,52%
SP6	46,55%	44,83%	48,28%	48,28%	56,90%	61,21%
SP7	45,45%	40,91%	42,42%	43,94%	34,85%	27,27%
SP8	61,02%	49,15%	52,54%	42,37%	49,15%	33,90%
SP9	48,57%	42,86%	50,00%	45,71%	52,86%	57,14%
SP10	48,24%	43,53%	50,59%	38,82%	60,00%	61,18%
MEAN	**47,24%**	**43,09%**	**46,03%**	**44,49%**	**48,87%**	**50,06%**

Table 4 presents a comparison of the results obtained from the two databases (SAVEE and IEMOCAP). It shows for the two contexts that the best performing wavelet feature bank is ERB, using entropy features, even if for the SAVEE database, the same accuracy is obtained from the energy and entropy features. We can also remark that SAVEE database is more performing in terms of classification than the IEMOCAP. The accuracy obtained with SAVEE is in the range [67.5%; 78.5%] when results with IEMOCAP are in the range of [43.09%; 50.06%]. This can be explained by the fact that SAVEE database is fully acted when IEMOCAP database contain spontaneous sessions. SAVEE is a fully acted database, where the recordings are controlled by the actors, and the sentences are phonetically balanced. The expression of emotion is clearer. IEMOCAP is an hybrid database, where speakers plays scripted scenarios (55% of the corpus) and spontaneous sessions (45% of the corpus). Moreover, we note that most of the sentences contain portions of silence, during which only low background noise can be heard. The accuracy difference obtained between the two databases can also be explained by the difference in terms of data sets. The SAVEE database contains 240 utterances vs. 2218 for IEMOCAP. As a general comment, recordings environment impact significantly levels of classification accuracies.

Table 4:Comparison of accuracy of databases

			SAVEE	IEMOCAP
SCALE	**MEL**	**Energy**	70.00%	47.24%
		Entropy	68.75%	43.09%
	BARK	**Energy**	68.75%	46.03%
		Entropy	67.50%	44.49%
	ERB	**Energy**	78.75%	48.87%
		Entropy	78.75%	50.06%

4. Conclusion

In this paper, we set up a speech emotion recognition system based on the wavelet packet energy and entropy features. To test the effectiveness of the system, we performed it on two sets of data SAVEE and IEMOCAP, one is full acted and other is scripted with spontaneous recording. Feature extraction was carried out using wavelet packet by partitioning the frequency axis analogous to the Mel, Bark and ERB scale. The results show that wavelet packet filter bank with ERB scale give promising classification accuracy for both of databases.

In the future works, it is advised to use others classification methods such Support Vector Machines (SVM) or Neural Networks. Moreover, it is proposed to add different noises to data to test the robustness of the proposed system. Combining emotional signals with psychological perception can be a promising idea to ameliorate the effectiveness of the system.

References

1. A.Tawari and M.M.Trivedi, “Speech Emotion Analysis: Exploring the Role of Context”,IEEE Transactions on Multimedia,vol 12,No 6,October 2010.
2. A.E.Ayadi, M.S.Kamel and F.Karray, “survey on speech emotion recognition:features, classifiation, schemes and databases”,Pattern Recognition, Vol 44, Issue 3,pp. 572-587, March 2011.
3. Z.Zeng, M.Pantic, GI. Roisman and T.S. Huang, “A survey of affect recognition methods: audio, visual, and spontaneous expressions”,IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol 31, No 1,January 2009.
4. C.Busso, M.Bulut, and S.Narayanan, “Toward effective automatic recognition systems of emotion in speech”, In Social emotions in nature and artifact: emotions in human and human-computer interaction. S. Marsella J. Gratch, Ed, Oxford University Press,2012.
5. D.Le and E.M.Provost, “Emotion recognition from spontaneous speech using hidden markov models with deep belief networks”, Automatic Speech Recognition and Understanding (ASRU). Olomouc, Czech Republic, December 2013.

6. C.Busso, A.Metallinou and S.S.Narayanan," Iterative feature normalization for emotional speech detection", IEEE Transactions on Affective Computing, Vol 4 , No 4 , pp.386 – 397, October 2013.
7. N.S. Fulmare, P.Chakrabarti and D.Yadav," Understanding and estimation of emotional expression using acoustic analysis of natural speech", International Journal on Natural Language Computing (IJNLC), Vol 2, No.4, October 2013.
8. F.Dellaert, T.Polzin and A. Waibel," Recognizing emotion in speech", Fourth International Conference on Spoken Language (ICSLP), pp. 1970 – 1973, Vol.3, October 1996.
9. B.Schuller, G.Rigoll, and M.Lang," Hidden markov model-based speech emotion recognition", IEEE International Conference on Acoustics, Speech, and Signal Processing, (ICASSP '03), Vol 2, April 2003.
10. O.W.Kwon,K.Chan,J.Hao and T.W.Lee" Emotion recognition by speech signals", 8th European Conference on Speech Communication and Technology, September 2003.
11. C.M.Lee and S.Narayanan," Emotion recognition using a data-driven fuzzy inference system", 8th European Conference on Speech Communication and Technology, September 2003.
12. N.Trivedi, V.Kumar, S.Singh, S.Ahuja and R.Chadha, "Speech Recognition by Wavelet Analysis", International Journal of Computer Applications,Vol 15, No 8, February 2011.
13. S. Mallat, "A Wavelet Tour of Signal Processing", New York, Academic Press, 1999.
14. C.Burrus and R.Gopinath, H. Guo, "Introduction to wavelets and wavelet transforms : a primer", Prentice Hall Upper Saddle River, 1998.
15. O.Farooq and S.Datta, "Mel filter-like admissible wavelet packet struture for speech recognition", IEEE signal processing letters,Vol 8, No 7, July 2001.
16. R. Narayanam,H. Dajani and S.Again, "Wavelet filter banks modeling of human auditory system for robust speech enhancement", Interntional Journal of Scientific and Engineering Research, Vol 3 , Issue 4, April 2012.
17. A.Biswas, P.K.Sahu,M.Chandra, "Admissible wavelet packet features based on human inner ear frequencyresponse for hindi consonant recognition", Computers and Electrical Engineering, 2014.
18. T.Vogt, E.Andre and J.Wagner "Automatic Recognition of Emotions from Speech: A Review of the Literature and Recommendations for Practical Realisation" , Affect and Emotion in Human-Computer Interaction, pp.75-91, 2008.
19. S. Haq, P. Jackson, and J. Edge, "Audio-visual feature selection and reduction for emotion classification", In Proc AVSP, pp. 185-190,2008.
20. C. Busso, M. Bulut, C. Lee, A.Kazemzadeh, E. Mower, S. Kim, J. Chang, S. Lee, and S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database", Journal of Language Resources and Evaluation, vol. 42, no. 4, pp. 335-359, December 2008..
21. S. J. Young, G. Evermann, M. J. F. Gales, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev and P. C. Woodland, The HTK Book, version 3.4, 2006.