International Journal of Applied Engineering Research
ISSN 0973-4562 Volume 10, Number 10 (2015) pp. 26543-26558
© Research India Publications
http://www.ripublication.com

Higher Order Mel-Frequency Cepstral and Autoregressive Reflection Coefficients in Recognizing Three Dimensions of Speech Emotions

A. Milton[a], S. Tamil Selvi[b]
[a]Department of Electronics and Communication Engineering, St. Xavier's Catholic College of Engineering, Chunkankadai, Nagercoil 629003, Tamil Nadu, India. e-mail: milton@sxcce.edu.in, Tel.: 91-9442602309.
[b]Department of Electronics and Communication Engineering, National Engineering College, Kovilpatti 628503, Tamil Nadu, India. e-mail: stsece@nec.edu.in

Abstract

Emotion recognition from speech signals is an essential research area in the field of machine intelligence that is gaining momentum in recent years. Successful sensing of emotional states of a person will make a machine think and take decisions like humans. More and more researches have been carried out to recognize emotions from speech signals by investigating different features, different classification algorithms and different types of databases. Further research is needed to implement emotion recognition system with high percentage of accuracy in real life applications. Mel-frequency cepstral coefficients (MFCC) and autoregressive reflection coefficients (ARRC) are the two spectral features most successfully used in the field of emotion recognition from speech signals. But all the research works that have used the above two features are confined to the use of lower order features. In this research, we investigate the usefulness of higher order MFCCs and ARRCs in recognizing the three dimensions of emotions namely activation, valence and control. Experimental results show that higher order MFCCs and ARRCs recognize emotions better than lower order coefficients. In order to improve the statistical significance of the results, we have used two databases. Speaker-independent train-test model classification is used to avoid the issue of over-fitting.

Keywords Speech emotion recognition, Mel frequency cepstral coefficients, Autoregressive reflection coefficients, Emotions in three dimensional space.

Introduction
Emotions in humans are basically provoked by evaluation of the actual or anticipated reactions of others or a situation [1,2]. Emotional states of humans are expressed through changes in speech, facial expression, posture and physiological processes. Different emotional states are associated with different physiological and psychological states. These physiological and psychological states result in the production of speech signals that carry emotional information [3]. Therefore, emotional state of a person can be identified by analyzing the speech signal of the person. The most important advantage of recognition of emotions from speech signal is that speech is the primary communication mode for humans. Acquisition of speech signals does not require physical contact with the speaker, even speech signal obtained through telephone can be used to recognize the emotional state. Emotion recognition from speech signals finds its application in many fields like human-computer interface, medical diagnosis, forensic voice analysis, emotional speech synthesis [4-6], computer game [7,8], computer-based tutorial system [9], call center [10,11] and on-line speech emotion recognition [12].

Speech emotion recognition system consists of different phases. The three major phases of automatic recognition of emotions from speech signals are i) acquisition and validation of emotional speech database, ii) extraction of speech features and iii) emotion classification. Database acquisition is the process of obtaining speech signals with different emotions and validation is a process by which the speech signals are validated by human perception test. Validation authenticates the emotional class of the speech signals and helps to retain the speech signals of good emotional quality. Speech is time-varying signal and it consists of heaps of information. The direct use of raw speech signal to recognize emotions is not suitable. Therefore, different features are described to represent different characteristics of speech signals. Feature extraction is the process of extracting desired features from speech signals. The extracted features will represent the emotional speech signals in the process of classification. In classification, the speech features with known emotional class are used to train a classifier. The trained classifier is used to predict the emotional class of test speech signals.

Emotions in Dimensional Space
In emotion classification problem, the emotions are considered as discrete categories by most of the researchers. Anger, boredom, disgust, fear, happiness, sadness and neutral are considered as the seven basic discrete emotions [13]. The basic emotions represent the different states of the brain system to deal with different types of situations [14]. A few researchers have considered emotions as regions in a two- or three-dimensional continuous space [13,15,16].

In two-dimensional representation, emotions are located by arousal and valence coordinates. Most of the basic emotions can be differentiated using arousal and valence dimensions [17-19]. The arousal dimension, also known as activation, ranges from low to high, and the valence dimension ranges from negative to positive. The arousal dimension describes the internal excitement of the speaker, low value represents calm and high value represents very active. The valence dimension describes the pleasantness of the stimuli, the positive valence indicates pleasantness and the negative valence indicates unpleasantness [20], i.e. valence indicates the speaker's positive or negative evaluation of a situation [3].

In three-dimensional representation, the emotions are plotted in a three-dimensional space by adding control (attention-rejection) dimension to the arousal and valence dimensions [21]. The necessity for the third dimension occurred because some emotions, such as fear and anger are very close in arousal and valence dimensions. Therefore, it is difficult to differentiate such emotions in two-dimensional space. This problem can be solved by adding a third dimension. Cowie and Cornelius [14] call the third dimension as perceived control which ranges from negative to positive. It is positive for anger and negative for fear. Koolagudi and Rao [22] refer to the third dimension as dominance or power. The third dimension is also known as stance; its positive value represents the advance approachable emotions, and the negative value represents the retreat approachable emotions. That is, it describes the degree of control that the speaker tries to take on the circumstances. Table 1 shows the different regional positions of seven basic emotions in three-dimensional space.

Table 1: Regional position of different emotions in three-dimensional space

Emotion	**Three-dimensional coordinates**		
	Activation	**Valence**	**Control**
Anger	HA	NV	PC
Boredom	LA	NV	PC
Disgust	LA	NV	NC
Fear	HA	NV	NC
Happiness	HA	PV	PC
Sadness	LA	NV	NC
Surprise	HA	PV	PC

HA – High activation, LA – Low activation, PV – Positive valence, NV – Negative valence, PC – Positive Control and NC – Negative control.

Wide varieties of speech features have been used by different researchers to recognize speech emotions. Considerable recognition improvement has been obtained in activation dimension unlike valence and control. Most of the features used in the field of emotion recognition from speech signals provide excellent classification of high activation emotions against low activation emotions. They have not been used in a successful way to provide the same degree of classification of positive valence emotions against negative valence emotions and positive control emotions against negative control emotions. Recognition of emotions in valence and control dimensions can be improved by a new speech features or by analyzing the existing features in a curious way. In this paper, we analyze the three-dimensional (activation-valence-control) emotion recognition performance of higher order Mel-frequency cepstral coefficients (MFCC) and autoregressive reflection coefficients (ARRC) against the traditional way of using only the lower order coefficients.

MFCCs and ARRCs in Speech Emotion Recognition

MFCCs are the most resourceful features ever used in the field of speech emotion recognition. Different researchers have examined MFCCs of different orders. In literature, the order of coefficient used by most of the researchers is 12^{th} order MFCCs [23-30]. Other investigated coefficient orders are 10^{th} [31], 20^{th} [25] and 24^{th} [32] order MFCCs. It is clear from the above literature review that higher order MFCCs are not utilized to recognize emotions from speech signals.

ARRCs are also useful features to recognize speech emotions. A very few researchers like Milton and Selvi [32] have examined ARRCs in the field of emotions recognition from speech signals. The only one examined order of ARRCs is 25. Examination of ARRCs beyond the order 25, may throw some more light on the usefulness of higher order ARRCs.

In this paper, we are motivated to investigate the emotion recognition performance of 12^{th}, 24^{th}, 36^{th}, 48^{th}, 60^{th} and 72^{nd} order MFCCs and ARRCs. We analyze the variation of classification performance with increasing order of coefficients and the effectiveness of higher order coefficients in recognizing emotions in each of the three dimensions. Our experimental results highlight the usefulness of higher order MFCC and ARRC features.

The rest of the paper is organized as follows: Section 2 details the emotional databases, extraction of MFCCs, extraction of ARRCs, support vector machine classifier (SVM) and experimental analysis. Section 3 provides the experimental results and discussion. Section 4 concludes the paper.

Analysis of Higher Order MFCCs and ARRCs

MFCCs and ARRCs in different order are extracted from emotional speech signals of two different databases namely Berlin emotional database and Enterface emotional database. The extracted features of different orders are investigated for their three-dimensional emotion recognition performance using SVM classifier.

Berlin Database

The Berlin emotional database [33] is the well-known studio-recorded emotional database. It consists of utterances spoken in seven emotional states like anger, boredom, disgust, fear, happiness, sadness and neutral. A total of ten actors, five male and five female, have uttered ten emotionally neutral German sentences in the above seven emotions. The recorded sentences were validated by a perception test. Based on the perception test, speech signals with naturalness greater than or equal to 60% and recognition rate greater than or equal to 80% have been retained in the database. Berlin database consists of 535 emotional speech samples with a minimum length of 1.5 seconds and a maximum length of 4 seconds.

Enterface Database

The Enterface database [34] is an induced audiovisual emotional database. It consists of utterances spoken in six emotional states such as anger, disgust, fear, happiness, sadness and surprise. Forty three naive speakers from fourteen different countries have uttered five sentences in English for each emotion in office environment. Out of the total forty three speakers, thirty four were males and nine were females. It is called induced database because the speakers were induced into different emotions by telling different stories. Each of the stories was selected in such a way that it induced a particular emotion. The database was validated by two research experts. A total of 1287 audio files in wave format with lengths ranging from 1 second to 3 seconds are available in the database.

Extraction of MFCC Features

MFCCs are extracted based on human auditory perception of sound. MFCCs are extracted by dividing the speech spectrum into different overlapping bands using Mel-frequency scale. The Mel-frequency scale is given by Equation (1).

$$\mathrm{F} = 2595 \log_{10}\left(\frac{\mathrm{f}}{700} + 1\right) \tag{1}$$

where, *F* is the Mel-frequency, and *f* is the normal frequency. The bands are linearly spaced at low frequencies and logarithmically spaced at high frequencies. Discrete cosine transform of logarithm of power for each band is the MFCC. We have extracted MFFCs of different orders using the following procedure.

Step 1: 20 ms Hamming window is defined

Step 2: 20 ms frame of speech signal is obtained using the Hamming window

Step 3: For each frame the spectrum is calculated by discrete Fourier transform

Step 4: The whole frequency range is converted into Mel-frequency scale by Equation (1)

Step 5: In Mel-scale the whole range is divided into *N* number of bands using triangular window. Where, *N* defines the number (order) of coefficients. In Mel-scale all the bands will have the same bandwidth

Step 6: Each Mel-band is converted back into the normal frequency band. In normal frequency scale, lower bands will have low bandwidth and higher frequency bands will have higher bandwidth. The bands are linearly spaced at low frequencies and logarithmically spaced at high frequencies

Step 7: Logarithm of power of each band is calculated

Step 8: Discrete cosine transform of logarithm of power provides N^{th} order Mel-frequency cepstral coefficients.

Step 9: Hamming window is shifted by 10 ms to select the next speech frame and the procedure is repeated from Step 2 till the MFCCs of last frame is calculated

Extraction of ARRC Features

In autoregressive (AR) model, also known as all-pole model, the speech samples *x(n)* are related to the excitation signal *u(n)* by the following difference equation [35]

$$x(n) = \sum_{k=1}^{P} a_P(k)x(n-k) + Gu(n) \tag{2}$$

where, $a_P(k)$ are the linear prediction coefficients and, *G* is the gain parameter.

The corresponding transfer function is given by Equation (3). When the observed speech signal is available, in AR model analysis, the task is to estimate the coefficients $a_P(k)$ and the gain parameter *G*.

$$H(z) = \frac{G}{1+\sum_{k=1}^{P} a_P(k) z^{-k}} \tag{3}$$

where, *P* is order of AR model.

Parameters of AR model of order *P* are estimated using a recursive algorithm. Reflection coefficients are obtained as by-products in the recursive algorithm. Reflection coefficients of order *P* are the collection of linear prediction coefficients $a_i(i)$, $i=1,2\ldots P$. AR model of order *P*, models vocal tract as an acoustic tube with *P* sections, with each section having different cross sectional area. The i^{th} reflection coefficient indicates the ratio of cross sectional area of i^{th} and $i+1^{st}$ acoustic tubes [35]. ARRC features are found to perform better than linear prediction coefficients. Their performance is comparable to MFCCs, and significant improvement in emotion recognition performance is obtained when ARRC features are combined with MFCC features [32].

In order to extract ARRCs, the speech signals are divided into 20 ms Hamming windowed frames with a frame shift of 10 ms. From each frame, ARRCs of different orders are extracted using Levinson-Durbin's recursive algorithm [35].

Frame level MFCCs and ARRCs extracted from frames of a speech signal are converted into sentence level features by the functionals like mean, median, maximum, minimum, mean absolute deviation, range, inter-quartile range, standard deviation, skewness and kurtosis.

Support Vector Machine Classifier

The classifier which is used to investigate the performance of different orders of MFCCs and ARRCs is support vector machine. SVM classifies the input features based on the discriminative nature present in some of the features. In SVM classification, the input patterns are mapped onto a high dimensional feature space, and the mapping is done in such a way that the non-linearly separable input patterns become linearly separable patterns in the feature space. The goal in training SVM is to find the separating hyperplane with maximum margin. The margin is the positive distance of the closest data point from the separating hyperplane. The hyperplane with maximum margin is called optimum hyperplane. In SVM, the optimum hyperplane that separates the patterns of two classes is found by the principle of structural risk minimization. The concept of optimum hyperplane in SVM classifier is shown in Figure 1. The data points close to the hyperplane are called support vectors. These support vectors are crucial in SVM training because they are the training samples that define the optimum hyperplane.

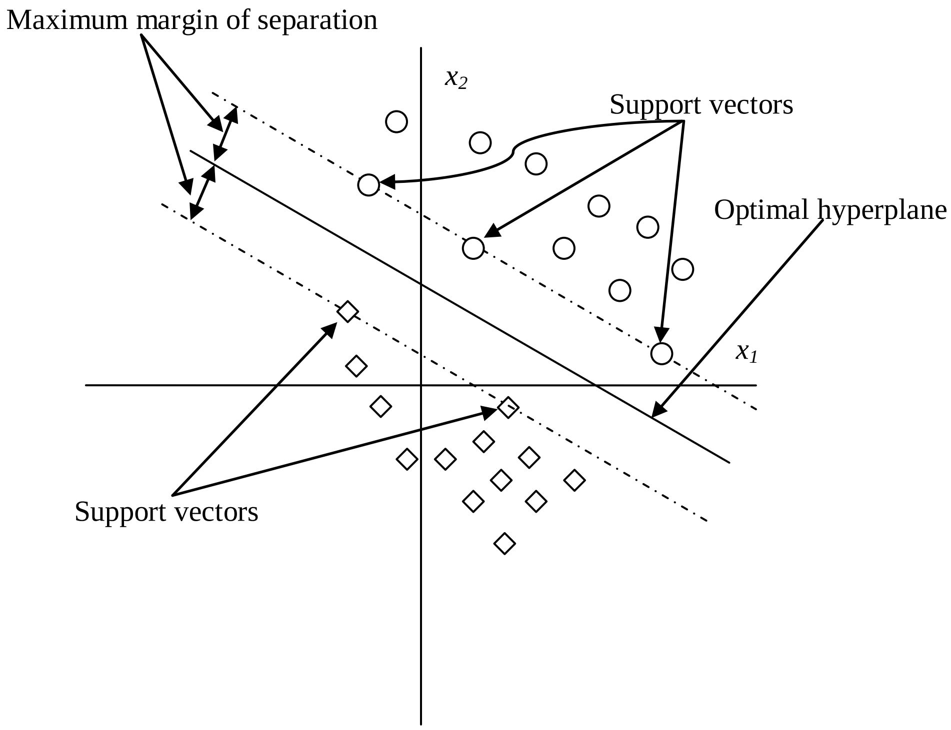


Figure 1: Illustration of the concept of SVM optimal hyperplane

We use quadratic kernel to transform the features into linearly separable high dimensional space. To find the optimum hyperplane, we use two types of methods, the first one is the least square method (LS) and the second one is the sequential minimal optimization (SMO) method.

Classification Analysis

In order to avoid the problem of over-fitting, each database is divided into two disjoint sets. One set acts as the training set and the other acts as the test set. The training and testing sets are divided so as to have speaker-independent classification system. In speaker-independent classification, the classifier is trained with features derived from speech signals uttered by a set of speakers and tested with features derived from speech signals uttered by another set of unknown speakers. In the case of Berlin database, the test set is formed by speech signals uttered by four speakers and in the case of Enterface database, the test set is formed by speech signals uttered by sixteen speakers. MFCC and ARRC features of different orders are extracted from the training set of each database as explained in features extraction sections. Training features with known emotional class are used to train SVM classifier. When features derived from test set are applied at the input of the classifier, based on the knowledge gained during training, it predicts the emotional class of the test set. Emotion classification performance of each order of MFCCs and ARRCs are analyzed separately with the LS and SMO SVM classifiers. Train-test experiments are conducted to investigate the classification of high activation emotions (anger, fear, happiness and surprise) against low activation emotions (boredom, disgust and sadness), positive valence emotions (happiness and surprise) against negative valence emotions (anger, boredom, disgust, fear and sadness) and positive control emotions (anger, boredom, happiness and surprise) against negative control emotions (disgust, fear and sadness).

Results and Discussion

Results obtained in different experiments are shown in Table 2. First column of Table 2 shows the speech features, second column shows the type of SVM classifier, third column shows the database, fourth column shows the binary emotional group in each of the three dimensions and fifth to tenth columns show classification accuracies of different order features. From the table, it is clear that the classification performance varies with speech features, classification algorithm, emotion dimension and type of database. In general, higher order MFCCs and ARRCs perform better than lower order coefficients in discriminating emotions in each of the three dimensions. We discuss below the performance of MFCCs and ARRCs with the help of scatter plots.

Performance of MFCCs

Classification performances of different orders of MFCC features in classifying emotions in each of the three dimensions for speech samples of Berlin and Enterface databases are shown in Figures 2 and 3. Figure 2 shows the performance of the features when least square SVM is used for classification and Figure 3 shows the performance when sequential minimal optimization SVM is used.

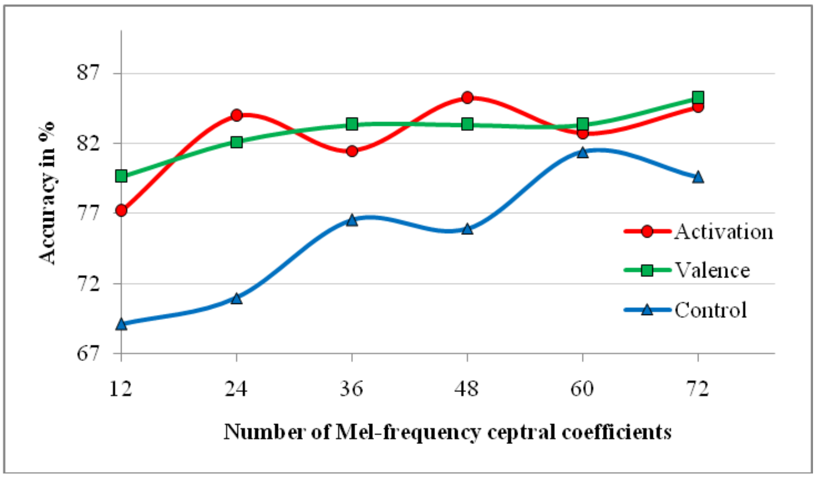
For the acted Berlin database, from Figure 2(a), it is clear that classification performance increases with increasing order of MFCCs for all the three dimensions. When the coefficient orders are lower, emotions in control dimension are classified with very low accuracy when compared with that of activation and valence dimensions. Only higher order MFCCs especially 60 and 72 classify emotions well in all the three dimensions. For the induced Enterface database, Figure 2(b) shows a steady fluctuating trend. Still the higher order coefficients especially 72 are able to differentiate emotions in activation, valence and control dimensions with high average recognition rate like the lower order of 24.

Table 2: Classification performance of MFCC and ARRC features of different orders in three-dimensional emotion space (Bold face indicates maximum accuracy for a database in recognizing a particular emotional dimension)

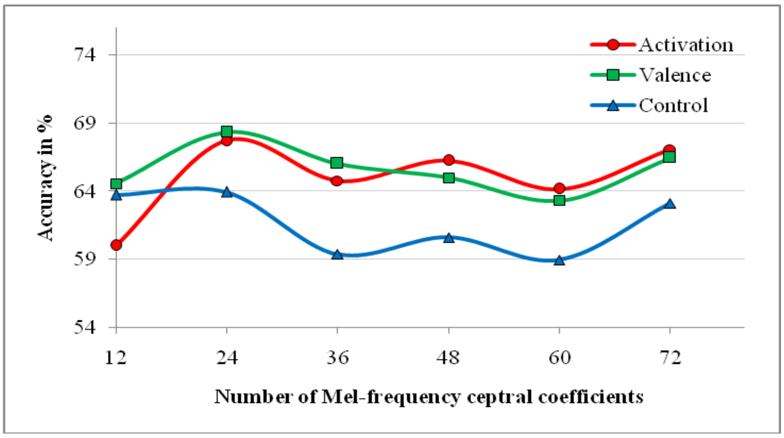
Features	SVM	Data-base	Order of coefficients	12	24	36	48	60	72
				Accuracy in %					
MFCCs	**LS**	**Berlin**	HA against LA	77.16	83.95	81.48	**85.19**	82.72	84.57
			PV against NV	79.63	82.10	83.33	83.33	83.33	85.19
			PC against NC	69.14	70.99	76.54	75.93	**81.40**	79.63
		Enter-	HA against LA	60.00	**67.71**	64.79	66.25	64.17	67.08

		face	PV against NV	64.58	68.33	66.04	65.00	63.33	66.46
			PC against NC	63.75	63.96	59.38	60.63	58.96	63.13
	SMO	**Berlin**	HA against LA	81.48	83.95	80.25	**85.19**	84.57	83.95
			PV against NV	81.48	83.33	83.33	83.33	82.72	**85.8**
			PC against NC	70.99	68.52	72.22	74.07	79.01	80.86
		Enter-face	HA against LA	61.46	67.08	65.42	66.88	64.79	67.29
			PV against NV	66.25	**69.17**	67.08	66.88	62.92	65.42
			PC against NC	**66.88**	63.13	60.42	61.25	59.58	63.33
ARRCs	**LS**	**Berlin**	HA against LA	74.69	77.78	83.33	82.10	79.01	80.25
			PV against NV	74.69	82.10	80.86	81.48	80.86	82.72
			PC against NC	61.11	71.60	72.84	65.43	64.20	60.49
		Enter-face	HA against LA	60.83	65.00	67.71	69.79	70.21	**72.08**
			PV against NV	51.25	55.21	58.33	55.63	61.25	**63.54**
			PC against NC	56.46	57.08	62.50	64.38	65.21	64.79
	SMO	**Berlin**	HA against LA	77.78	78.40	**84.57**	**84.57**	80.86	80.25
			PV against NV	77.78	81.48	81.48	82.10	82.10	**83.33**
			PC against NC	63.58	**73.46**	72.22	67.28	62.96	62.35
		Enter-face	HA against LA	68.33	66.25	68.54	69.38	71.04	**72.08**
			PV against NV	54.17	55.83	57.92	57.92	62.08	63.33
			PC against NC	56.25	60.00	63.96	64.38	**66.46**	65.21

MFCCs – Mel-frequency cepstral coefficients, ARRCs – Autoregressive reflection coefficients, LS – Least square method, SMO – Sequential minimal optimization method, HA – High activation, LA – Low activation, PV – Positive valence, NV – Negative valence, PC – Positive control and NC – Negative control.



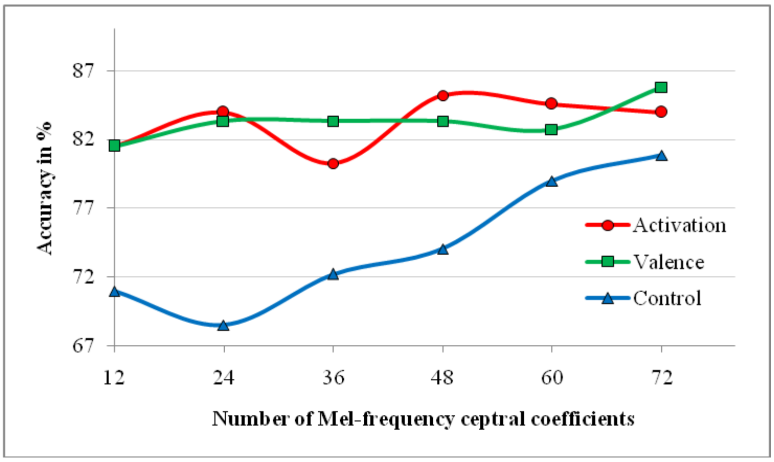
(a)



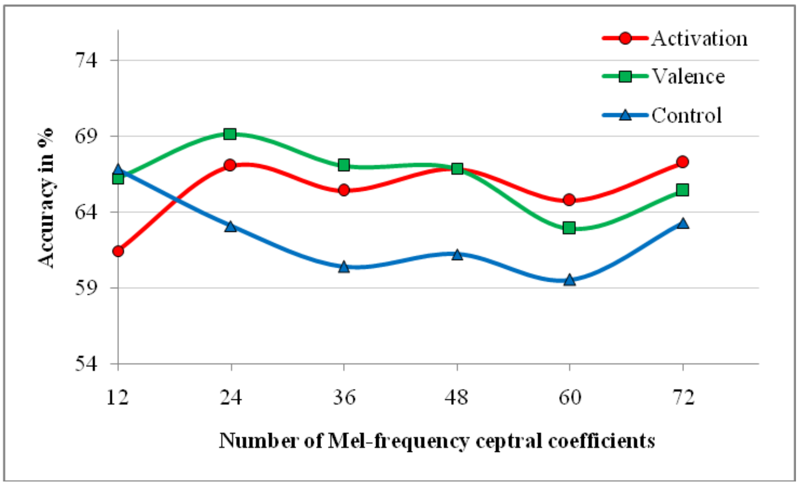
(b)

Figure 2: Classification performance of different orders of MFCCs with least square SVM in classifying emotions in three dimensions for (a) Berlin and (b) Enterface database

When MFCC features are classified with SMO SVM, for Berlin database, Figure 3(a) shows a clear increasing classification performance with increasing coefficient orders. Emotions in control dimension are classified with low accuracy but higher order coefficients considerably improve classification accuracy to high value. Only higher order coefficients are able to differentiate emotions well in all the three dimensions. Figure 3(b) shows the performance of SMO SVM with Enterface database. It shows a steady fluctuating trend with coefficient orders. Emotions in control dimension are classified better by the lowest order coefficients also appreciably good by the highest order coefficients. Middle order coefficients fail to differentiate emotions in control dimension. Still the higher order coefficient especially 72 are able to differentiate emotions in activation, valence and control dimension with high average recognition rate like the lower and middle orders 24 and 48.



(a)



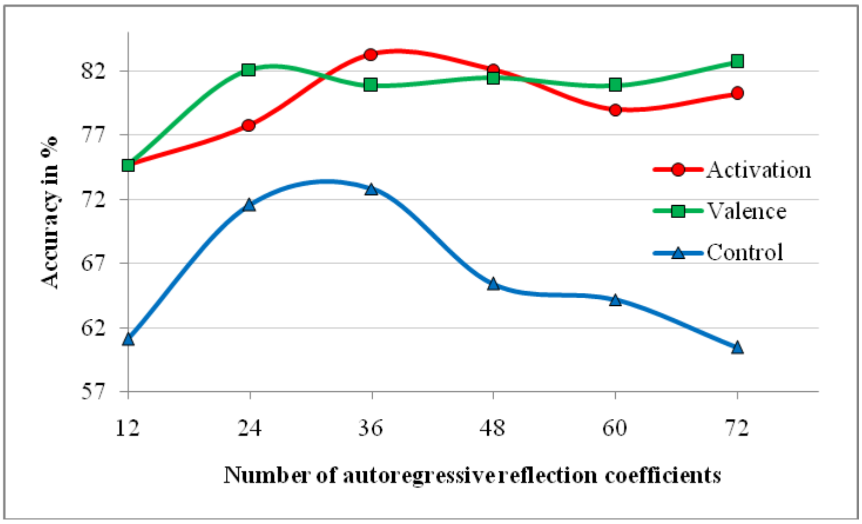
(b)

Figure 3: Classification performance of different orders of MFCCs with sequential minimal optimization SVM in classifying emotions in three dimensions for (a) Berlin and (b) Enterface database

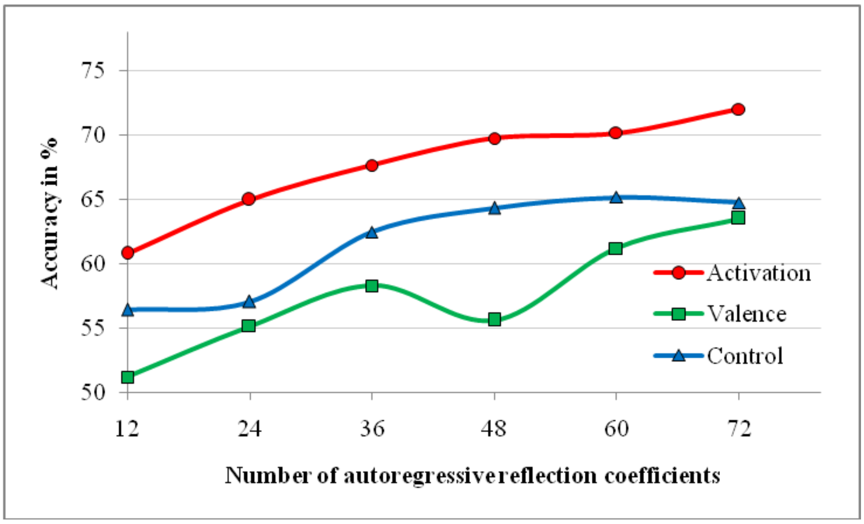
In general, for acted database, the performance increases with increasing coefficient orders, and for induced database even though the increasing trend is missing, higher order MFCCs differentiate emotions well in all the three dimensions. Examination of higher order coefficients beyond 72 may throw some more light on the usefulness of higher order coefficients. Classification of emotions in control dimension shows significantly increasing trend with higher order coefficients. Experimental results of this research emphasize the fact that MFCCs carry more emotional information when the features are derived by dividing the frequency range into more number of bands with reduced bandwidth.

Performance of ARRCs

Classification performances of ARRCs derived from Berlin database, with LS SVM are shown in Figure 4(a). Classification of emotions in activation and valence dimensions show increasing performance with increasing ARRC order. Classification performance of positive control emotions against negative control emotions decreases with increasing order except a moderate increase with middle order ARRCs. Emotions in all the three dimensions are classified well with middle order ARRCs of order 36. It is evident from Figure 4(b) that classification of emotions of Enterface database in activation, valence and control dimensions show a clear increasing performance with increasing ARRC orders except a slight decrease in control dimension in middle order. ARRCs classify emotions in control dimension better than emotions in valence dimension, a trend that is not exhibited with MFCCs.



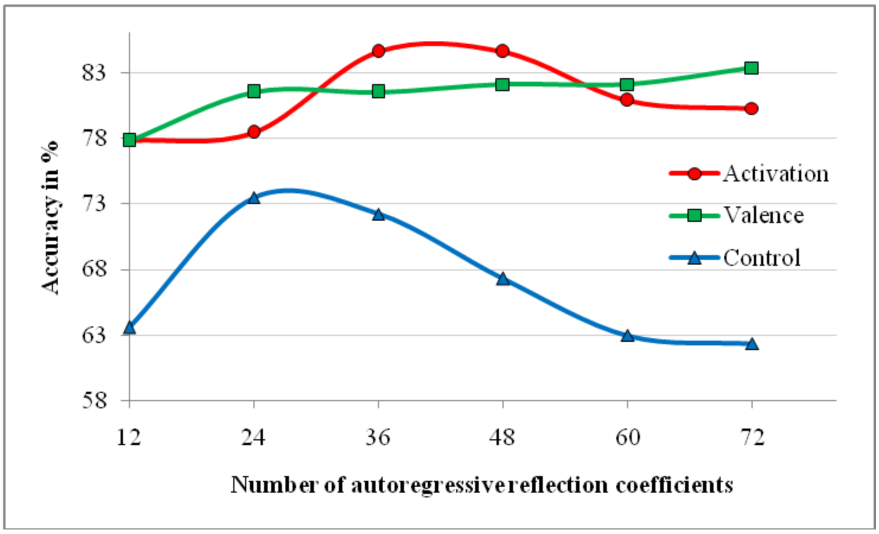
(a)



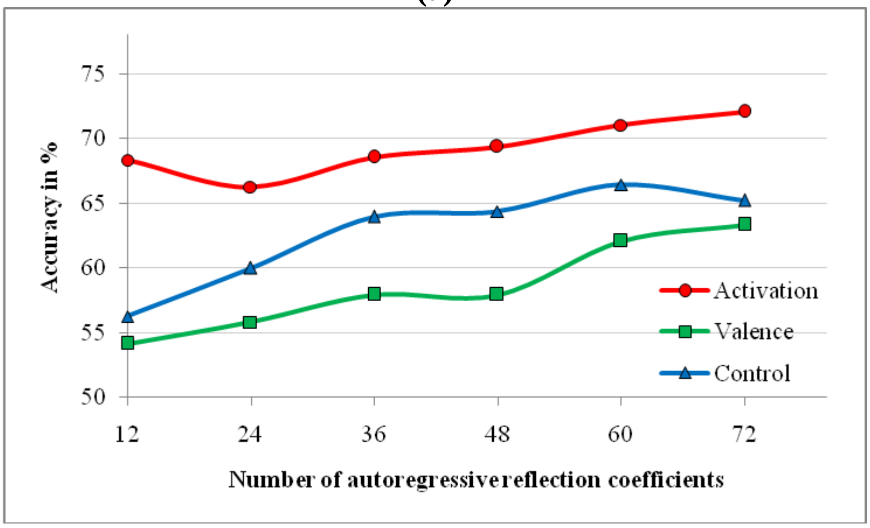
(b)

Figure 4: Classification performance of different orders of ARRCs with least square SVM in classifying emotions in three dimensions for (a) Berlin and (b) Enterface database

Classification analysis of ARRC features derived from Berlin and Enterface databases, with SMO SVM is shown in Figure 5. Figure 5(a) shows the classification performance for Berlin database. Middle order ARRCs recognize emotions better than lower and higher order ARRCs in all the three dimensions. Discrimination of emotions in control dimension decreases with increasing order. Figure 5(b) shows classification performance for Enterface database, emotion discrimination performance in all the three dimensions increases significantly with increasing order of ARRCs and emotions in control dimensions are differentiated with better accuracy than emotions in valence dimension.



(a)



(b)

Figure 5: Classification performance of different orders of MFCCs with sequential minimal optimization SVM in classifying emotions in three dimensions for (a) Berlin and (b) Enterface database

In general, irrespective of classification algorithm, in the case of acted database, emotions in activation and valence dimensions show increasing classification performance with increasing ARRC order. But the emotions in control dimension show a decreasing classification performance with increasing order of ARRCs. In the case of induced database, emotions in all the three dimensions show increasing classification performance with increasing order of ARRCs. It emphasize the fact that considering vocal tract as concatenation of more number of acoustic tubes provides more emotional information.

Conclusion

In this paper, we have investigated the three-dimensional emotion recognition performance of 12th, 24th, 35th, 48th, 60th and 72nd order Mel-frequency cepstral coefficients and autoregressive reflection coefficients. Speaker-independent classification analysis with support vector machine classifier shows that higher order coefficients recognize emotions better than the traditional lower order coefficients. They recognize emotions better than lower order coefficients irrespective of different classification algorithms and different types of databases. Only higher order coefficients recognize emotions better in all the three dimensions known as activation, valence and control.

Experimental results show that emotion recognition performance of Mel-frequency cepstral coefficients and autoregressive reflection coefficients increases with increasing order of coefficients. Therefore, investigation beyond the order 72 may help to find the optimum order.

References

[1] Scherer, K. R., Schorr, A., and Johnstone, T., 2001, “Appraisal Processes in Emotion Theory, Methods, Research,” Oxford University Press, New York.

[2] Parkinson, B., 2008, “Emotions in Direct and Remote Social Interaction: Getting through the Spaces between Us,” Computers in Human Behavior, 24(4), pp. 1510-1529.

[3] Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., and Taylor, J. G., 2001, “Emotion Recognition in Human-Computer Interaction,” IEEE signal Processing Magazine, 18(1), pp. 32-80.

[4] Inanoglu, Z., and Young, S., 2009, “Data-Driven Emotion Conversion in Spoken English,” Speech Communication, 51(3), pp. 268-283.

[5] Murray, I. R., and Arnott, J. L., 2008, “Applying an Analysis of Acted Vocal Emotions to Improve the Simulation of Synthetic Speech,” Computer Speech and Language, 22(2), pp. 107-129.

[6] Tao, J., Kang, Y., and Li, A., 2006, “Prosody Conversion from Neutral Speech to Emotional Speech,” IEEE Transaction on Audio, Speech and Language Processing, 14(4), pp. 1145-1154.

[7] Klein, J., Moon, Y., and Picard, R. W., 2002, “This Computer Responds to User Frustration: Theory, Design and Results,” Interacting with Computers, 14(2), pp. 119-140.

[8] Scheirer, J., Fernandez, R., Klein, J., and Picard, R. W., 2002, “Frustrating the User on Purpose: a Step toward Building an Effective Computer,” Interacting with Computers, 14(2), pp. 93-118.

[9] Ai, H., Litman, D. J., Forbes-Riley, K., Rotaru, M., Tetreault, A., and Purandare, A., 2006, “Using System and User Performance Features to

Improve Emotion Detection in Spoken Tutoring Systems," Proceedings of Interspeech 2006, pp. 797-800.
[10] Lee, C. M., Narayanan, S. S., and Pieraccini, R., 2002, "Combining Acoustic and Language Information for Emotion Recognition," Proceedings of ICSLP, pp. 873-876.
[11] Ang, J., Dhillon, R., Krupski, A., Shriberg, E., and Stolcke, A., 2002, "Prosody-based Automatic Detection of Annoyance and Frustration in Human-Computer Dialog," Proceedings of ICSLP, pp. 2037-2039.
[12] Wang, F., Sahli, H., Gao, J., Jiang, D., and Verhelst, W., 2014, "Relevance Units Machine based Dimensional and Continuous Speech Emotion Prediction," Multimedia Tools and Applications, DOI: 10.1007/s11042-014-2319-1.
[13] Yang, B., and Lugger, M., 2010, "Emotion Recognition from Speech Signals using New Harmony Features," Signal Processing, 90(5), pp. 1415-1423.
[14] Cowie, R. and Cornelius, R. R., 2003, "Describing the Emotional States that are Expressed in Speech," Speech Communication, 40(1-2), pp. 5-32.
[15] Grimm, M., Kroschel, K., and Narayanan, S., 2007, "Support Vector Regression for Automatic Recognition of Spontaneous Emotions in Speech," Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2007, Honolulu, Hawaii, USA, Vol. 4, pp. 1085-1088.
[16] Grimm, M., Kroschel, K., Mower, E., and Narayanan, S., 2007, "Primitive-based Evaluation and Estimation of Emotion in Speech," Speech Communication, 49(10-11), pp. 787-800.
[17] Kim, E. H., Hyun, K. H., Kim, S. H., and Kwak, Y. K., 2009, "Improved Emotion Recognition with a Novel Speaker-Independent Features," IEEE ASME Transaction on Mechatronics, 14(3), pp. 317-325.
[18] Wöllmer, M., Schuller, B., Eyben, F., and Rigoll, G., 2010, "Combining Long Short-term Memory and Dynamic Bayesian Networks for Incremental Emotion-Sensitive Artificial Listening," IEEE journal of Selected Topics in Signal Processing, 4(5), pp. 867-881
[19] Tarng, W., Chen, Y. Y., Li, C. L., Hsie, K. R., and Chen, M., 2010, "Applications of Support Vector Machines on Smart Phone Systems for Emotional Speech Recognition," World Academy of Science, Engineering and Technology, 72, pp. 106.113.
[20] Caponetti, L., Buscicchio, C. A., and Castellano, G., 2011, "Biologically Inspired Emotion Recognition from Speech," EURASIP Journal on Advances in Signal Processing, Vol. 2011-24.
[21] Schlosberg, H., 1954, "Three Dimensions of Emotions," Psychological Review, 61(2), pp. 81-88.
[22] Koolagudi, S. G., and Rao, K. G., 2012, "Emotion Recognition from Speech using Source, System, and Prosodic Features," International Journal of Speech Technology, 15(2), pp. 265-289.

[23] Schuller, B., Müller, R., Lang, M., and Rigoll, G., 2005, "Speaker Independent Emotion Recognition by Early Fusion of Acoustic and Linguistic Features within Ensembles," Proceedings of Interspeech 2005, Lisbon, Portugal, pp. 805-808.

[24] Yoon, W. J., and Park, K. S., 2011, "Building Robust Emotion Recognition System on Heterogeneous Speech Databases," IEEE Transaction on Consumer Electronics, 57(2), pp. 747-750.

[25] Koolagudi, S. G., and Rao, S. K., 2011, "Two Stage Emotion Recognition based on Speaking Rate," International Journal of Speech Technology, 14(1), pp. 35-48.

[26] Bozkurt, E., Erzin, E., Erdem, C. E., and Erdem, A. T., 2011, "Formant Position based Weighted Spectral Features for Emotion Recognition," Speech Communication, 53(9-10), pp. 1186-1197.

[27] Sheikhan, M., Bejani, M. and Gharavian, D., 2013, "Modular Neural-SVM Scheme for Speech Emotion Recognition using ANOVA Feature Selection Method," Neural Computing & Applications, 23(1), pp. 215-227.

[28] Albornoz, E. M., Milone, D. H., and Rufiner, H. L., 2011, "Spoken Emotion Recognition using Hierarchical Classifiers," Computer Speech and Language, 25(3), pp. 556-570.

[29] López-Cózar, R., Silovsky, J., and Kroul, M., 2011, "Enhancement of Emotion Detection in Spoken Dialog Systems by Combining Several Information Sources," Speech Communication, 53(9-10), pp. 1210-1228.

[30] Wu, C. H., and Liang, W. B., 2011, "Emotion Recognition of Affective Speech based on Multiple Classifiers using Acoustic-Prosodic Information and Semantic Labels," IEEE Transaction on Affective Computing, 2(1), pp. 10-21.

[31] Shami, M., and Verhelst, W., 2007, "An Evaluation of the Robustness of Existing Supervised Machine Learning Approaches to the Classification of Emotions in Speech," Speech Communication, 49(3), pp. 201-212.

[32] Milton, A., and Selvi, S. T., 2014, "Class-Specific Multiple Classifiers Scheme to Recognize Emotions from Speech Signals," Computer Speech and Language, 28(3), pp. 727-742.

[33] Burkhardt, F., Paeschke, A., Rolfes, M., Senlmeier, W., and Weiss, B., 2005, "A Database of German Emotional Speech," Proceedings of the Interspeech 2005, Lisbon, Portugal, pp. 1517-1520.

[34] Martin, O., Kotsia, I., Macq, B., and Pitas, I., 2006, "The Enterface'05 Audio-Visual Emotion Database," Proceedings of IEEE Workshop on Multimedia Database Management, Atlanta, pp. 1-8.

[35] Makhoul, J., 1975, "Linear Prediction: a Tutorial Review," Proceedings of the IEEE, 63(4), pp. 561-580.