CHARACTERIZATION OF EMOTIONS USING THE DYNAMICS OF PROSODIC FEATURES

K Sreenivasa Rao, Ramu Reddy, Sudhamay Maity and Shashidhar G Koolagudi

School of Information Technology
Indian Institute of Technology Kharagpur
Kharagpur - 721302, West Bengal, India.
E-mail: ksrao@iitkgp.ac.in, {ramu.csc, friendsudha, koolagudi}@gmail.com

ABSTRACT

In this paper the dynamics of prosodic parameters are explored for recognizing the emotions from speech. The dynamics of prosodic parameters refer to local or fine variations in prosodic parameters with respect to time. The proposed dynamic features of prosody are represented by : (1) sequence of durations of syllables in the utterance (duration contour), (2) sequence of fundamental frequency values (pitch contour) and (3) sequence of frame energy values (energy contour). Indian Institute of Technology Kharagpur Simulated Emotion Speech Corpus (IITKGP-SESC) is used for analyzing the proposed prosodic features for recognizing the emotions [1]. The emotions considered in this work are anger, disgust, fear, happiness neutral and sadness. Support vector machines (SVM) are explored to discriminate the emotions using the proposed prosodic features. Emotion recognition performance is analyzed separately, using duration patterns of the sequence of syllables, pitch contours and energy contours, and their recognition performance is observed to be 64%, 67% and 53% respectively. Fusion techniques are explored at feature and score levels. The performance of the fusion-based emotion recognition systems is observed to be 69% and 74% for feature and score level fusions,respectively.

1. INTRODUCTION

Human beings use emotions extensively to convey their intensions and feelings. For an effective human machine interaction, the machine should be able to adapt its interaction policies according to user's emotional state. Hence the sophisticated interface to the machine should ensure understanding of emotions expressed by humans and responding back with appropriate emotions. The capability of emotional analysis by machine has several applications. For example, in call center applications, the machine first analyzes the emotional state of the customer, and responds to the customer by itself, if he/she is in positive mood, otherwise transfers the call to human attendant [2]. The automatic way of analyzing the emotions in speech is useful for indexing and retrieving the audio files based on emotions [3].

The emotion specific characteristics of the speech can be attributed to the characteristics of vocal tract system and excitation source, behavioral characters of the speaker, suprasegmental characteristics and linguistic information. Emotion specific vocal tract information is mainly represented by spectral features like Mel-frequency cepstral coefficients (MFCC), linear prediction cepstral coefficients (LPCC) and their derivatives [4]. The parameters like pitch, duration and energy are used as basic prosodic features, and their derivatives extracted from longer speech segments are used to categorize the emotions present in the speech [3]. Most of the existing emotion recognition systems are developed using spectral features, prosodic features and their combination [3, 4, 5, 6, 7].

Dellert *et al* analyzed F_0 information for emotion classification and reported that minimum, maximum and median values of F_0, and slopes of F_0 contours are the salient features [5]. Williams C.E. *et al* proposed LFPCs (Log frequency power coefficients) for classifying six emotions [6]. Ververidies *et al* used short time suprasegmental features and their statistics such as pitch frequency (F_0), energy, formant locations and their bandwidths for analyzing the emotions [3]. Gobl *et al* combined vocal tract features, voice overtones in various types of speech and pitch dynamics to classify the emotions [7]. In addition to pitch related information, Kwon *et al* used log energy, formants, mel based energy, MFCC's, Δ's, $\Delta\Delta$'s for classifying the emotions [7]. Wang *et al* used 55 features (25-prosodic, 24-MFCC's and 6-formant frequencies) for recognizing six emotions [7].

Most of the existing work used Prosodic features mainly by exploiting their gross statistics at the utterance level such as maximum, minimum, mean, standard deviation and slope are exploited in the literature, but the variation in prosodic features with respect to time are not explored. It is known that the dynamics of the prosodic parameters (i.e., local or fine variations in prosodic parameters with respect to time) are also crucial, along with the gross prosodic parameters in manifesting the emotions at the suprasegmental level. Hence, in this work we are exploring the time varying prosodic parameters for characterizing the emotions present in speech. In this study, the durations of the sequence of syllables present in the utterance (duration contour), sequence of fundamental frequency (pitch (F_0)) values (pitch contour) and the sequence of frame energies (energy contour) are considered as dynamics of the prosodic parameters. Support vector machine (SVM) models are explored for discriminating the emotions using the above mentioned features. The reason for choosing the SVM model for this task is that, it will provide better discrimination even with fewer number of examples, and also with the feature vectors of higher dimensions. In this work, capturing dynamics of the prosody, requires high dimensional feature vector. The number of speech utterances in the database is also limited. Hence, we have opted SVM model for discriminating the emotions using the dynamics of the prosodic parameters.

The paper is organized as follows: In Section 2, the motivation for choosing the proposed features (dynamics of prosody) for recognizing the emotions is discussed. Description of the emotional speech database used in this work is given in Section 3. Recognition performance of the developed SVM models is given in Section 4. Summary, conclusions of the present work, and the scope of future work are discussed in the last section.

2. MOTIVATION

Our basic intuition is that the duration patterns of the sequence of syllables, intonation and intensity patterns (pitch and energy contours) are specific to each emotion, and with that we can able to discriminate different emotions. For demonstrating this subjective listening tests are conducted on the speech utterances, where the dynamics of the prosody, specific to the emotion are removed. Average emotion recognition results computed from the listening tests are given in Table 1. From the results, it is observed that listeners are not able to recognize the emotions, when the dynamics of prosody, specific to the emotion are removed. The recognition performance is observed to be very bad, if the dynamics of all prosodic parameters are removed together.

3. DATABASE

Indian Institute of Technology Kharagpur Simulated Emotion Speech Corpus (IITKGP-SESC) is recorded using 10 (5 male and 5 female) professional artists from All India Radio (AIR) Vijayawada, India [1]. 15 Telugu sentences are used for recording. All the sentences have emotionally neutral in meaning. Each of the artists has to speak the 15 sentences in 8 basic emotions in one session. The number of sessions considered for preparing the database is 10. The total number of utterances in the database is 12000 (15 sentences $\times 8$ emotions $\times 10$ artists $\times 10$ sessions). The eight basic emotions present in the database are: Anger, Sadness, Disgust, Fear, Happyness, Neutral, Sarcastic and Surprise. The speech signal is sampled at 16 kHz, and represented as 16 bit numbers. In this study, we have analyzed the emotional speech of a single female speaker for six emotions.

Table 1. Emotion recognition performance using subjective listening tests for (i) original emotional sentences (S1), (ii) emotional sentences after replacing the dynamics in syllable durations by their average duration (S2), (iii) emotional sentences after replacing the dynamics in the sequence of pitch values by their average pitch value (S3), (iv) emotional sentences after replacing the dynamics in the sequence of frame energies with their average frame energy (S4), and (v) emotional sentences after replacing the dynamics in durations of syllables, sequence of pitch values and sequence of frame energies together with their respective average values (S5).

Emotion	Emotion recognition performance (%)				
class	S1	S2	S3	S4	S5
Anger	81	23	17	73	23
Disgust	72	32	22	65	22
Fear	69	27	19	62	26
Happy	78	39	43	72	30
Neutral	83	62	95	75	21
Sad	82	65	62	73	25

4. EXPERIMENTAL RESULTS AND DISCUSSIONS

In this work, 900 utterances (150 utterances × 6 emotions) from IITKGP-SESC are considered for analyzing the time varying local variations of prosodic parameters for discriminating the emotions. 120 utterances of each emotion are used for developing the models, and 30 utterances per emotion are used for validation. For developing the SVM model for the specific emotion, feature vectors derived from the utterances of desired emotion are used as positive examples, and the feature vectors derived from the utterances of all other emotions (other than the desired emotion) are used as negative examples. The block diagram of the Emotion Recognition (ER) system using SVM models is shown in Fig. 1. For evaluating the performance of the ER system, the feature vectors derived from the test utterances are given as input to all SVM models. The output of the each model is given to decision logic. Decision ligic determines the emotion category based on the highest score among the six evidences provided by the emotion models. In this work, first we analyzed the capability of individual prosodic parameters for discriminating the emotions. Later, we have explored the feature and score level fusion techniques for

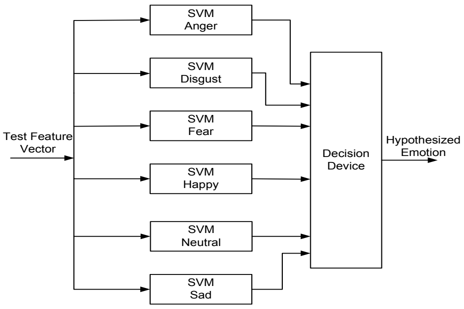

Fig. 1. Emotion recognition system using SVM models.

improving the performance. Three ER systems developed using individual prosodic features are:

1. ERS-1: ER system using the sequence of durations of syllables within the utterance.
2. ERS-2: ER system using the sequence of fundamental frequency (F_0) values (pitch values) of the utterance.
3. ERS-3: ER system using the sequence of frame level energies of the utterance.

Durations of syllables are derived by using ergodig hidden Markov models. Sequence of syllable durations form the feature vectors for representing the duration contours. Since, the number of syllables in different utterances is not constant, the size of feature vectors is not same. But for developing SVM classifiers, all feature vectors should have the same size. In view of this, the dimension of the feature vectors is kept fixed to the maximum number of syllables present in the whole set of utterances used in this work. For the utterances containing the lower number of syllables, compensation of length is done by zero padding. For representing the intonation pattern of an utterance, the sequence of pitch values of all the voiced frames of the utterance is used. For deriving the fixed dimensional feature vector resampling is performed. Similarly for representing energy contour, sequence of voiced frame energies followed by resampling is used. Pitch and energy values are derived from speech signal, using a frame size of 20 msec and a frame shift of 10 msec. For representing the durations of the sequence of syllables, a 20-demensional feature vector is used. For representing the sequence of pitch values (pitch contour) and energy values (energy contour), feature vectors of different sizes are explored for recognizing the emotions. The optimal size of the feature vector to capture the variations in pitch and energy contours for recognizing the emotions is observed to be 100.

The performance of the ER systems developed by using individual prosodic features is given in Table 2. Columns 2-7, 8-13 and 14-19 show the performance of the ER systems developed by using duration contour, pitch contour and energy contour, respectively. The diagonal entries of the corresponding ER systems indicate the correct recognition performance of the emotions considered in this study. The other entries indicate the percentage of misclassification. The average recognition performance of the ER systems developed using duration, pitch and energy contours is observed to be 64%, 67% and 53% respectively. From the classification results, it is observed that anger, disgust, fear and happy form a group (see rows 1-4 of Table 2), and other two emotions neutral and sad form the other group (see rows 5 and 6 of Table 2). Hence, the classification and misclassification is observed within the respective groups. This phenomenon is also observed in the performance of ER systems developed by feature and score level fusion methods (see Table 3). Since, Telugu language has distinctive phonological durations for the sound units for different emotions, hence the emotion recognition using duration contour seems to be better compared to energy contour.

For enhancing the performance of the individual ER systems, fusion techniques are tried out at feature and score levels. In this work, feature level fusion is performed by concatenating the individual prosodic features, and the ER system is developed using the concatenated feature vectors. The performance of the ER system developed using feature level fusion is given in columns 2-7 of Table 3. The average recognition performance is observed to be 69%.

Table 3. Performance of the emotion recognition systems developed by using (1) Feature level fusion and (2) Score level fusion. The entries in the table indicate the percentage of recognition.

	Feature level fusion						Score level fusion					
	A	D	F	H	N	S	A	D	F	H	N	S
A	80	0	3	17	0	0	77	0	3	20	0	0
D	0	57	23	13	7	0	10	63	17	3	7	0
F	0	20	63	7	10	0	7	17	70	3	0	0
H	23	0	3	67	7	0	23	0	0	70	7	0
N	0	7	0	0	73	20	0	0	0	7	83	10
S	0	7	7	0	13	73	0	0	0	3	17	80

In this work, score level fusion is performed by summing the weighted confidence scores (evidences) derived from the ER systems developed using individual prosodic features. The weighting rule for combining scores of individual modalities is as follows: $c^m = \frac{1}{m}\sum_{i=1}^{m} w_i c_i$, where c^m is the multimodal confidence score, w_i and c_i are weighting factor and confidence score of the i^{th} modality, and m indicates number of modalities used for combining the scores. In this work, we have combined three modalities: (1) Model developed using durational features, (2) Model developed using sequence of pitch values and (3) Model developed using sequence of frame energies. In our study, one of the weights (w_i) is varied in steps of 0.1 from 0 to 1, and the other weights are determined using the formula: $w_j = \frac{1-w_i}{m-1}$, where j =1 to m and $j \neq i$, i = 1 to m. In this

Table 2. Performance of the emotion recognition systems developed using the features representing (1) Duration contour, (2) Pitch contour and (3) Energy contour. The entries in the table indicate the percentage of recognition. (A: Anger, D: Disgust F: Fear, H: Happy, N:Neutral and S: Sad)

	Duration contour						Pitch contour						Energy contour					
	A	D	F	H	N	S	A	D	F	H	N	S	A	D	F	H	N	S
A	77	0	7	13	3	0	54	0	13	20	13	0	40	20	13	20	7	0
D	10	50	23	7	10	0	20	60	20	0	0	0	13	34	20	13	0	20
F	3	33	47	10	7	0	13	7	67	0	0	13	0	7	80	0	0	13
H	27	6	10	57	0	0	17	0	0	83	0	0	7	26	0	54	13	0
N	0	0	7	3	73	17	7	13	0	0	67	13	0	13	0	7	54	26
S	0	0	3	7	10	80	7	0	13	0	26	54	0	0	13	0	33	54

study, weighting factors associated to each system is varied from 0 to 1, with the steps of 0.1. With this we get a total of 33 combinations (11 combinations with respect to each weighting factor) of weighting factors. The recognition performance of the combined system for various combinations of the weighting factors is shown in Fig. 2. It is observed that the best recognition performance is about 74% for the weighting factors 0.2, 0.6 and 0.2 for the confidence scores of duration, pitch and energy contour based ER systems respectively. The details of the recognition performance are shown in columns 8-13 of Table 3.

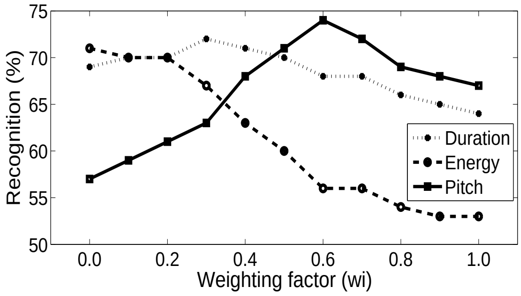

Fig. 2. Recognition performance for different combinations of weighting factors.

5. SUMMARY AND CONCLUSIONS

The basic goal of the paper was to investigate whether the dynamics of the prosodic parameters at the utterance level, contain emotion specific information. The prosodic parameters considered in this study were syllable level durations, sequence of fundamental frequency values, and sequence of frame level energy values. SVMs were used as classification models for developing different ER systems. The performance of the ER systems developed by durations of syllables, pitch contours and energy contours are observed to be 64% 67% and 53% respectively. Fusion techniques were explored by combining the features and scores of the individual systems. The performance of the ER system was improved by combining the scores, rather than combining the features. The recognition performance of the system using score level fusion was observed to be 74%. The recognition performance may be further improved by combining the prosodic features derived from the word, syllable and phrase levels. The present work canbe extended to analyze the ER performance for speaker independent case. One can explore the combination of static and dynamic features of prosody for discriminating the emotions.

6. REFERENCES

[1] S. G. Koolagudi, S. Maity, V. A. Kumar, S. Chakrabarti, and K. S. Rao, *IITKGP-SESC : Speech Database for Emotion Analysis*, issn: 1865-0929 ed., ser. Communications in Computer and Information Science. Springer, August 2009, vol. 40.

[2] C. M. Lee and S. S. Narayanan, “Toward detecting emotions in spoken dialogs,” *IEEE Trans. Speech and Audio Processing*, vol. 13, no. 2, pp. 293–303, March 2005.

[3] D. Ververidis and C. Kotropoulos, “A state of the art review on emotional speech databases,” in *Eleventh Australasian International Conference on Speech Science and Technology*, Auckland, New Zealand, Dec. 2006.

[4] Y.Wang and L.Guan, “An investigation of speech-based human emotion recognition.” IEEE 6th Workshop on Multimedia Signal Processing, 2004, pp. 15–18.

[5] F. Dellert, T. Polzin, and A. Waibel, “Recognizing emotion in speech.” Philadelphia, PA, USA: 4th International Conference on Spoken Language Processing, October 3-6 1996.

[6] T. L. Nwe, S. W. Foo, and L. C. D. Silva, “Speech emotion recognition using hidden Markov models,” *Speech Communication*, vol. 41, no. 4, pp. 603–623, Nov. 2003.

[7] A. I. Iliev, M. S. Scordilis, J. P. Papa, and A. X. Falco, “Spoken emotion recognition through optimum-path forest classification using glottal features,” *CSL*, vol. in press, 2009.