Prosody Classification: Confirmation to Prosody Conversion

Priyanka B. Dhawale[1] Pallavi S. Deshpande[2]
[1,2]Department of Electronics and Telecommunication
[1,2]Smt. Kashibai Navale COE, Pune, India

Abstract— Prosody classification is one of the challenging tasks in prosody conversion. To determine whether we can convert the prosody or not, prosody classifications is main task. If we can classify the prosody, then it's sure we can convert the prosody. In this paper, we focused on the classification of prosody. We demonstrated the approach of classifying the prosody using the multi-class support vector machine. In first part, we considered spectral as well as prosodic features for classification. In second part, we considered only prosodic features. In second part, the classification accuracy is more. In SVM, we used the combination of RBF (radial basis function) and polynomial kernel function which gave better accuracy than single RBF kernel function. The successful prosody classification leads to confirmation of prosody classification.

Key words: Prosody classification, prosody, prosody conversion, support vector machine, kernel function, RBF, polynomial

I. INTRODUCTION

Important speaker identification related information is carried out in prosody. Prosody is not related to only the content spoken in the sentence. But prosody also depends on the speaker's speaking style [2]. If we see, for same example different people speak in different style. Also same person can speak same sentence differently, if we ask him to speak several times [2]. In general, it is difficult to define the term prosody. But in general prosody is the pitch or energy related term [3]. Prosody is the rhythmic distribution of phonemes. The automatic prosody classification systems mainly depend on the choice of correct features and use of appropriate classifier. There are various classification techniques including Gaussian Mixture Models (GMM), Hidden Markov Models (HMM), K-nearest neighbors (KNN), artificial neural networks (ANN), and (SVM) support vector machines etc. have been developed in the speech processing. In these classifiers, SVM has shown to provide a better generalization performance in different pattern recognition problems than traditional techniques [1].

II. DATA COLLECTION AND PRE-PROCESSING

for the accurate classification of prosody, the speech database should be more qualitative. The database should be more balanced prosodic so as to classify the prosody. The trained person can only do this. So we recorded the speech database from the professional speakers in clean environment. The samples were recorded using noise cancelling microphones. We used the Marathi speech database of our own. In this database, we recorded 320 samples consists of neutral, interrogative and exclamatory prosody. Data pre-processing stage consists of data sampling, data framing and windowing:

A. Sampling:

In sampling process, we convert the continuous time signal into the discrete time signal. Fictitious switch is the efficient way to show the sampling process. The switch is kept closed for a small interval of time T, at that time the output is generated. If T second is the time span between the successive samples, then sampling frequency is given by [6],

$$F = 1/t \text{ Hz.} \quad \ldots\ldots\ldots\ldots (1)$$

B. Framing:

Framing is the process of separating the speech samples into the small number of frames which is having length of 10 to 20 msec. In the system, the voice signal is divided into the frames of N samples. We keep the distance between adjacent frames as M to overcome the issue of overlapping and the shifting between frames is kept as 10 samples [6].

C. Windowing:

Now the next stage is to application of window to each individual frame. It helps in minimizing the discontinuities in signal at the beginning and end of the each frame [6].

III. FEATURES EXTRACTION AND MAPPING FUNCTION

Spectral envelope and basic frequency is considered by researchers to convert the emotion [4, 12, 15].

Fundamental Frequency (Pitch):

Fundamental frequency is the relative maximum and minimum tone received at the ear. Vocal cords produce the number of vibrations per second. Pitch depends on the number of vibrations per seconds. We used the autocorrelation method [9].

Energy:

The loudness is represented by the energy. The energy is calculated for each segment in frame by summing all the squared values of the samples amplitudes [10].

Pitch difference and energy difference:

As usual the pitch and energy difference is the between the neighboring samples. More the variation which may show the active emotions.

Formants (frequency and bandwidth for the first five formants, thus eight features):

Shape of the vocal tract determines the formants and they are inflamated by the various emotions. E.g. loudness results in higher mean value of the first formant frequency in all vowels.

MFCC (Mel Frequency Cepstrum Coefficients):

MFCC shows the state of emotion in the short calculations [1]. If the speech frequency is low then MFCC shows the good frequency resolution. MFCC shows the phonetic properties of the speech.

IV. PROPOSED PROSODY CLASSIFICATION METHOD

The proposed method consists of three parts as speech analysis [7, 14], training and testing phase. Speech analysis consists of speech data pre-processing. After data pre-processing the required feature vector is prepared which is used to train the support vector machine. In our approach we prepared two types of feature sets. In first set, we considered the features as energy, energy difference, formant, formant bandwidth, pitch difference, MFCC and duration. And in

second set we removed the formant, formant bandwidth, MFCC features [11]. We used two set of feature vector to identify the prosodic feature and confirm it for prosody conversion [8, 13]. So in first is not that good. So tested the SVM using second feature vector set, at this time accuracy is better than the previous one. In the first stage testing, we used the RBF kernel function. To increase the more accuracy, we used the hybrid kernel function. The hybrid kernel function consists of the RBF and polynomial kernel function. The SVM trained with hybrid kernel function gave the better results than the RBF kernel function. Stage we trained SVM using first feature set, but the accuracy

In training phase, first the data pre-processing is performed on the speech samples. Then the required feature vector is prepared. In MATLAB code, we add all feature analysis code in one function and saved the calculated feature vector in one excel. So that we can directly use excel for feature vector. Then the support vector machine is trained with the feature vectors.

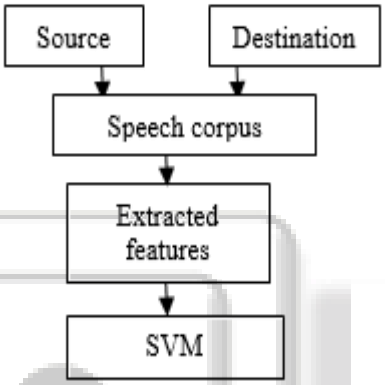


Fig. 1: Training phase of prosody classification system.

In testing phase, we used the testing data samples. Then the feature vector is prepared for the same testing data samples. The feature vector is given to the trained SVM to test the data.

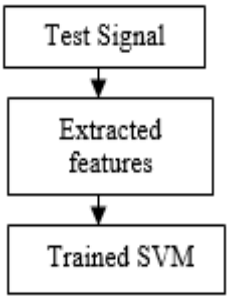


Fig. 2: Testing phase of prosody classification system.

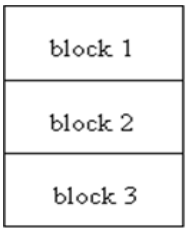
V. RESULTS

Accuracy achieved by the Gaussian RBF hybrid kernel is higher than the Gaussian RBF kernel function. Below table shows actual results where we can clearly suggest the better approach by Gaussian RBF hybrid kernel since it gives the better accuracy compare to linear RBF approach.

kernels	Classification-accuracy interrogative	Classification-accuracy Neutral
Linear	65%	72%
Gaussian	87%	75%
	RBF	RBF

Table 1: Accuracy using different kernel functions.

Objective Evaluation (K-fold cross validation):Here we have 60 sentences to be classified .we Split randomly the data in K=3 groups with roughly the same size Taking turns using one group as test set and the other k-I as training samples.



	train	test
Run 1	1,2	3
Run 2	1,3	2
Run 3	2,3	1

Fig. 3: fold cross validation

Following table shows the efficiency of table:

	Classification-accuracy % Linear RBF	Classification-accuracy % Gaussian RBF	
Run 1	70	76	Neutral
	65	83	Interrogative
Run 2	79	72	Neutral
	70	89	Interrogative
Run 3	67	77	Neutral
	60	89	Interrogative

Table 2: 3fold cross validation

Following waveform shows the neutral, interrogative and exclamatory prosody for Marathi sentence "Shree shalela jato":

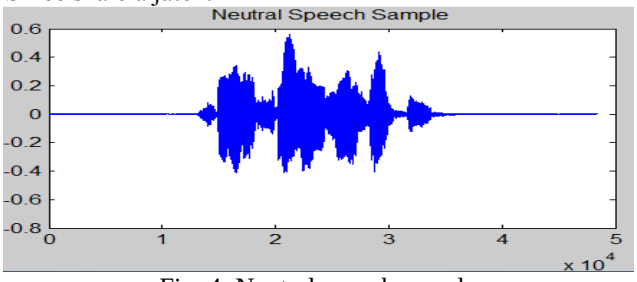


Fig. 4: Neutral speech sample

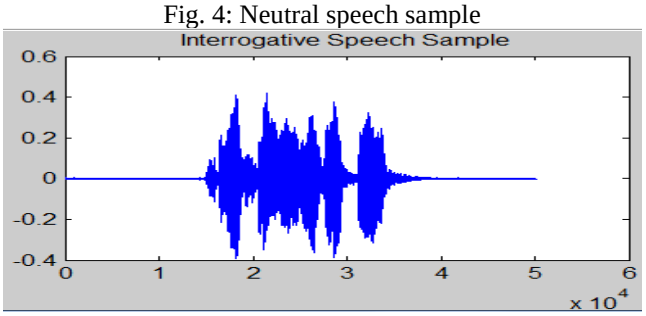


Fig. 5: Interrogative speech sample

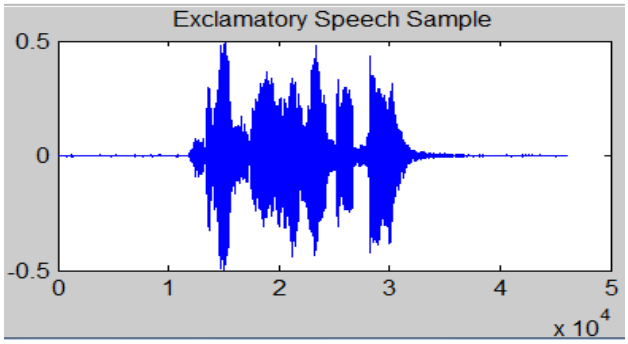


Fig. 6: Exclamatory speech sample

VI. CONCLUSION

This paper is first step in performing the prosody conversion. Voice conversion consists of the two main parts as spectral and prosody transformation. But the prosody transformation is the most challenging research. The first step in it is to confirm whether we can convert prosody using the speech features. So in this paper, we classified prosody according to the feature sets and we came to know the features required for prosody transformation. We successfully classified the prosody using second feature set. So this work leads to first step in prosody conversion.

REFERENCES

[1] Pooja Yadav and Gaurav Agarwal, " Speech Emotion classification using machine learning", International journal of computer applications (0975-8887) Volume 118-No.13 May 2015.
[2] Elina E. Helander, "A novel method for prosody prediction in voice conversion", IEEE 2007.
[3] Anderson F.Machado and Marcelo Queiroz, " Voice conversion : A critical survey", open – access article distributed under the terms of the creative commons attribution licence 2010.
[4] Ryo Aihara, Ryoichi Takashima, Tetsuya Takiguchi and Yasuo Ariki, " GMM-based Emotional voice conversion using spectrum and prosody features", American journal of signal processing 2012, 2(5): 134-138.
[5] Ms.C.D.Pophale and Prof.J.S.Chitode, "Embedding prosody into neutral speech", IOSR
[6] Mrs.R.B.Shinde and Dr. V.P.Pawar, "Dynamic Time Warping using MATLAB and PRAAT" ,International journal of scientific and engineering research, Volume 5, issue 5, May 2014. ISSN 2229-5518.
[7] Vishnu Mohan, "Analysis and synthesis of speech using matlab", International journal of advancements in Research and technology, Volume 2, Issue 5, May-2013.
[8] Jianhua Tao, Yongguo Kang and Aijun Li, "Prosody conversion from neutral speech to emotional speech", IEEE transactions on audio, speech and language processing, Vol.14,No.4, July2006.
[9] Jagannath Nirmal, Pramod Kachare, Suparva Patnaik and Mukesh Zaveri, "Cepstrum liftering based voice conversion using RBF and GMM", International conference on communication and signal processing, April 3-5, 2013, India.
[10] C A Manjare and S D Shirbahadurkar," Speech modification for prosody conversion in expressive marathi text-to-speech synthesis", 2014 International conference on signal processing and integrated networks (SPIN),IEEE 2014.
[11] Hajer Rahali, Zied Hajaiej and Noureddine Ellouze, "A Comparative study: Gammachirp wavelets and auditory filter using prosodic features of speech recognition in noisy environment", International Journal of computer science and security (IJCSS), Volume (8):Issue (2):2014.
[12] Yannis Stylianou, "Voice transformation: A survey", IEEE 2009.
[13] Divya setia, Maninder Singh Suri ans Anurag Jain, "Emotion conversion in Hindi Language", Proceedings of the 4th national conference;INDIACom-2010 computing for nation development, February 25-26,2010.
[14] Yegnanarayana B, Veldhuis R.N.J, "Extraction of vocal-tract system characteristics from speech signals", IEEE transactions on speech and audio processing, Vol. 6, 313-327, July 4, 1998.
[15] Wu, Y. H., Lin, S. J., & Yang, D. L. (2013, September). A mobile emotion recognition system based on speech signals and facial images. In Computer Science and Engineering Conference (ICSEC), 2013 International (pp. 212-217). IEEE.

