Importance of Non-Uniform Prosody Modification for Speech Recognition in Emotion Conditions

V. V. Vidyadhara Raju, Hari Krishna Vydana, Suryakanth V Gangashetty, and Anil Kumar Vuppala
Speech and Vision Lab, Kohli Center on Intelligent Systems, IIIT Hyderabad
vishnu.raju@research.iiit.ac.in, hari.vydana@research.iiit.ac.in, svg@iiit.ac.in, anil.vuppala@iiit.ac.in

***Abstract*—A mismatch in training and operating environments causes a performance degradation in speech recognition systems (ASR). One major reason for this mismatch is due to the presence of expressive (emotive) speech in operational environments. Emotions in speech majorly inflict the changes in the prosody parameters of pitch, duration and energy. This work is aimed at improving the performance of speech recognition systems in the presence of emotive speech. This work focuses on improving the speech recognition performance without disturbing the existing ASR system. The prosody modification of pitch, duration and energy is achieved by tuning the modification factors values for the relative differences between the neutral and emotional data sets. The neutral version of emotive speech is generated using uniform and non-uniform prosody modification methods for speech recognition. During the study, IITKGP-SESC corpus is used for building the ASR system. The speech recognition system for the emotions (anger, happy and compassion) is evaluated. An improvement in the performance of ASR is observed when the prosody modified emotive utterance is used for speech recognition in place of original emotive utterance. An average improvement around 5% in accuracy is observed due to the use of non-uniform prosody modification methods.**

***Index Terms*—prosody, automatic speech recognition, IITKGP-SESC, uniform prosody modification, non-uniform prosody modification.**

I. Introduction

Automatic speech recognition (ASR) is the process of converting the speech signal into sequence of symbols. A degradation in the performance of practical ASR system is observed due to mismatch in training and testing environments. Speech recognition in various emotional environments is a crucial aspect to be addressed in human-machine interaction. Emotion in speech majorly reflect the mental state of speaker and inflicts the changes in the physiological aspects of respiration, phonation and articulation [1]. The manifestation of these physiological changes majorly attribute to change in prosody parameters such as pitch, duration and energy [2].

Prosody modification involves the process of manipulating the pitch and duration of the speech without introducing the spectral and temporal distortions in it [3] [4]. In the literature, different techniques have been proposed for prosody modification [5] [6] [7]. These prosody modification techniques are classified into time domain and frequency domain approaches [8]. The influence of different emotions on various prosody parameters have been studied in [2] [9]. The relative change in prosody parameters have been studied at uniform and non-uniform levels. These relative changes in the prosody parameters are reported as prosody modification factors. There have been studies to quantify the relative changes in the prosody parameters as modification factors to generate an emotive utterance from neutral utterance and vice-versa. In [10], the prosody modification factors at uniform and non-uniform levels are studied for generating the neutral version of a emotive speech. The effectiveness of the modification factors of non-uniform levels are superior when compared to modification factors at uniform prosody levels. Most of the studies [11] [12] aim at generating the emotive version of a neutral utterance but there have been a very minimal attempts [13] to generate the neutral version of a emotional utterance. The motivation of the paper is to generate the neutral version of emotive speech and to evaluate the speech recognition system performance on the prosody modified speech [14]. Majority of above methods attempt to measure the effectiveness of prosody modification based on perceptual scores but applicability of these methods in the perspective of practical ASR system is explored in this paper.

In this study, uniform and non-uniform prosody modification method is explored for generating a neutral version of an emotive utterance to improve performance of ASR system. The remaining paper is organized as follows: Section II describes the details of the baseline experimental setup. In Section III, the details of the speech recognition in emotion conditions are discussed. Finally, Section IV gives the summary and scope for further studies.

II. Experimental setup for ASR system

This Section consists of two sub sections, A and B. Section II(A) provides the information about the emotion corpus used for the study. Section II(B) explains about the baseline ASR system performance.

A. Emotion Database

The database consisting of spontaneous natural emotions is required for the analysis of emotional speech. The emotion databases developed by different research groups are categorized as simulated, semi-natural and near to natural database. In this paper a simulated emotion corpus is prefered where the speech is recorded from professional speakers by prompting them to enact emotions through a given text in a language. There have been many standard simulated databases such as Danish emotion speech database (DES) and Berlin emotion

speech database (EMO-DB). The main reason for not considering these databases as there is no sufficient data to build an ASR system. These emotion corpus are limited only to 10 speakers. The Emotion database considered in this paper is Simulated Emotion Speech Corpus [15] which is collected by Indian Institute of Technology, Kharagpur (IITKGP-SESC). The database collected is for Telugu (Indian) language which consists of around 1500 utterances spoken by radio artists. The recordings collected from 10 professional speakers (5 male and 5 female) of 15 sentences each spoken in 8 emotions and repeating these in 10 separate sessions. The entire duration of the total database is seven hours. The ASR system is trained with neutral speech recorded from 40 speakers and the recordings from 10 speakers are used for testing in different emotion modes. In this paper the emotions considered other than neutral mode are happy, anger and compassion modes.

B. Baseline ASR system Performance

The baseline ASR system is built on Sphinx-3 tool kit developed by Carnegie Mellon University (CMU). The language model employed is a ARPA format trigram model built from CMUCLMTK toolkit. A word level HMM acoustic model is trained on the IITKGP-SESC corpus using the data from 40 speakers in neutral mode. This simulated emotion corpus consists of 60 words and testing is performed on 10 speakers. The speech recognition system is evaluated on the three different emotions of anger, happy and compassion.

The ASR system performance in terms of word accuracy for the neutral, anger, happy and compassion is shown in Table I. From the results in Table I, it can be observed that there is a degradation in the system performance observed for the three emotions of anger, happy and compassion. Column 2 in Table I describes the performance of the ASR system which is trained on neutral speech. The performance of the ASR system observed is better in case of neutral speech as the system is trained on the same neutral speech and a degradation in the performance is observed for speech in the other emotions (anger, happy and compassion). More degradation of the speech recognition system performance is observed in the case of happy emotion.

TABLE I. ASR system performance trained on neutral speech and tested on different emotions (neutral, anger, happy and compassion)

Emotion	ASR word accuracy (%)
Neutral	96.51
Anger	85.37
Happy	75.37
compassion	79.27

III. Speech Recognition in emotion conditions

In this paper, the effectiveness of non-uniform prosody modification is studied over uniform prosody modification method. The performances of these methods is compared with the study presented in [14] using FAST prosody modification method. The epoch based prosody modification is prefered where it employs the zero frequency filtering (ZFF) method to provide the accurate estimation of epochs [5] [6]. This section consists of three sub sections, A, B and C. Section III (A), III (B) and III (C) describes the ASR system in emotional environment using FAST, uniform and non-uniform prosody modification.

The block diagram for the proposed speech recognition system in emotional environments is presented in Fig.1. The emotive speech utterance considered is segmented to three different regions i.e. starting, middle and end regions. The prosody modification factors for the corresponding regions are chosen from Table.III. The prosody parameters of emotional speech are modified using non-uniform prosody modification and the neutral version of emotive speech is generated. This generated neutral speech is considered for speech recognition and the performance of ASR system is evaluated.

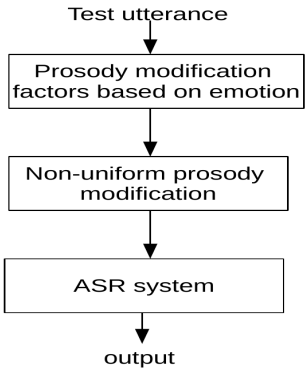

Fig. 1. Block diagram description for ASR system in emotional environments speech using non-uniform prosody modification

A. Performance of speech recognition system using FAST prosody method

In the FAST prosody modification [6] [16] the assumption is that the target neutral utterance is readily available to perform the prosody manipulation on the emotive utterance. Emotion in speech inflicts the changes the in physiological aspects. In emotion the majority of changes are observed in excitation source and prosody (pitch, duration and energy) parameters. The source excitation features are majorly reflected by the linear prediction (LP) residual. The three steps involved in the fast prosody manipulation is to derive the instants of significant excitation (epochs) from the speech signal by zero frequency filtering (ZFF) method [17], the next step is to derive a modified new epoch sequence according to the desired prosody parameter and the last step is to generate a modified speech signal from the modified epoch sequence. In this approach the duration of both emotive and neutral utterance is made equal using DTW alignment. LP residual from the neutral utterance is used to synthesize the neutral version of an emotive utterance and the synthesized neutral version is used for speech recognition instead of emotional utterance. The performance of the speech recognition system is shown in the column 2

of Table IV. The results obtained using FAST prosody method have shown the better improvement in the ASR system.

B. ASR system performance on uniform prosody modified speech

The FAST prosody method fails when there is no parallel neutral utterance readily available in the real-world scenario. The main aim of this work is to improve the speech recognition performance without disturbing the existing ASR system. The major challenge for the uniform prosody modification method lies in the generation of neutral version for the given emotive speech. The influence of the three emotions anger, happy and compassion on the prosody components of Pitch, duration and energy are studied. The relative changes in the prosody parameters for the emotive and neutral speech are identified and they are reported as the modification factors for the prosody components. The prosody modification factors to convert the emotive speech to neutral speech are considered from the analysis study reported in [18] [13]. In this method the relative differences of the prosody (pitch, duration and energy) parameters of emotional and neutral speech are considered. The average modification factors are considered from the analysis of different emotional databases such as EMO-DB, IITKGP-SESC and DES datasets of different languages. These modification factors are determined by tuning them for the differences between the trained neutral data and the emotional test sets. The prosody of the entire utterance is modified using a single modification factor on a phase vocoder. The modification factors used to convert the emotional utterance to neutral utterance are shown in Table II. The results obtained using uniform prosody modified speech is shown in column 4 of Table IV.

TABLE II. Ratio of prosody features of anger, happy, compassion emotions to neutral

	Pitch	Duration	Energy
Anger	0.59	0.62	0.72
Happy	0.86	0.92	1.01
Compassion	1.25	1.31	1.20

Table II indicates the uniform modification factors required to convert the anger, happy and compassion utterance to target neutral utterance. Column 2 in Table II indicates the mean fundamental frequency i.e. average pitch value, column 3 in Table II indicates the F_0 range. In this method the prosody parameters from columns 2, 4 and 5 from Table II are considered for converting the anger emotion to neutral utterances. These generated neutral version files are passed to the speech recognition and its performance is reported in column 4 of Table IV. The system performance is poor for happy and compassion emotions when compared to anger emotion.

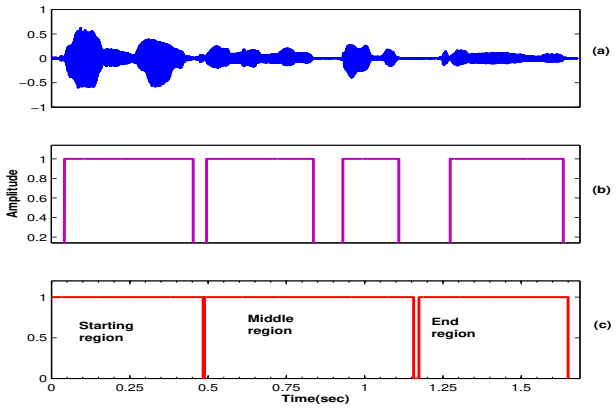

Fig. 2. Segmentation of emotional speech to perform non-uniform prosody modification where Fig.2(a) represents the emotional speech signal, Fig.2(b) represents the voice and unvoiced detection using ZFF signal and Fig.2(c) represents the segmentation of the emotional speech into starting, middle and end regions

C. ASR system performance for non-uniform prosody modified speech

In uniform prosody modification approach, it seems to be a crude way to have a single and fixed scaling factor per feature. There is a need to tune these modification factor values at a much finer level. The idea to perform the non-uniform prosody modification is considered from the study [13] by segementing the emotional speech utterance into three different regions. The prosody modification factors for different emotions of anger, happy and compassion have been reported in detail to convert them into a neutral utterance. The corresponding modification factors for each segment is used to generate the neutral version of emotive utterance and the generated neutral utterance is used for speech recognition. The tuned modification factor values required to perform non-uniform prosody modification are presented in Table III.

Anger Modification Factors			
Modification factors	Starting words of a sentence	Middle words of a sentence	Ending words of a sentence
Pitch	0.86	0.93	0.94
Duration	1.38	1.19	1.21
Energy	1.08	0.96	1.15
Happy Modification Factors			
Pitch	0.61	0.55	0.51
Duration	0.86	0.94	0.97
Energy	1.3	1.3	1.07
Compassion Modification Factors			
Pitch	0.72	0.65	0.54
Duration	0.90	1.04	1.1
Energy	1.4	1.1	0.75

TABLE III. Non-uniform prosody modification factors for converting the emotive utterance to neutral.

In Table III the modification factors for the prosody

TABLE IV. *ASR system performance for different prosody modification methods. Column 3 shows the results of the ASR trained directly with emotional speech. The evaluated results on FAST prosody method [14] is shown in Column 4, The results of Uniform prosody modification method is shown in Column 5 and Column 6 shows the results of Non-uniform prosody modification method*

	Word Accuracy in (%)				
Emotion	baseline	ASR trained with Emotional speech	FAST prosody [14]	Uniform prosody	Non-uniform prosody
Neutral	96.51	96.51	96.51	96.51	96.51
Anger	85.37	95.32	90.39	86.12	**89.03**
Happy	75.37	94.49	84.27	72.16	**80.29**
Compassion	79.27	92.29	86.46	75.37	**84.34**

parameters of pitch, duration and energy are provided. The modification factors are provided are based on position of segments in the emotional speech utterance. In this method, speech signal between consecutive voiced segments is considered as an acoustic word. The evidence about the voicing of the speech signal is computed using the approach presented in [19]. Based on the number of acoustic words the utterance is segmented to three regions i.e, starting middle and ending regions as shown in Fig.2. The prosody parameters of the emotional utterances are modified using these modification factors in Table III and the generated neutral utterance is used for speech recognition.

The effectiveness of the above three methods III (A), III (B), III (C) is studied for speech recognition system in emotional environments. The speech recognition system performance obtained by the three different prosody modification methods are reported in Table IV. The maximum ASR system performance is observed in the case when the system is directly trained with emotional speech. Due to the use of uniform modification factors a slight improvement has been observed in anger emotion. The uniform modification factors cannot effectively model the relative changes in prosody inflicted by emotion and a poor performance has been observed for the other emotions which is shown in column 5 of Table IV.

Though non-uniform method uses static modification factors, these modification factors are at much finer level in the speech utterance i.e. words and the performance of non-uniform method is superior to uniform method. In non-uniform method the relative changes are effectively captured and a better neutral version is generated giving a better word error rate which is shown in column 6 of Table IV.

IV. Summary and Scope for future work

A degradation is observed on the neutral speech ASR system when operated in emotional environments. Improvement in the system performance is shown without disturbing the existing neutral trained ASR system by exploring the prosody parameters at the pre-processing level. The neutral version generated from the emotional speech has yielded a better performance in the case of non-uniform prosody modification. This non-uniform prosody modified speech is used in the practical ASR application where the parallel neutral corpus is not readily available. The study can be extended to other emotions apart from the basic emotions of neutral, anger, happy and compassion. The other non-normal conditions apart from the emotional speech can be studied.

References

[1] A. Batliner, S. Steidl, D. Seppi, and B. Schuller, "Segmenting into adequate units for automatic recognition of emotion-related episodes: a speech-based approach," *Advances in Human-Computer Interaction*, vol. 1, p. 3, 2010.

[2] V. Mittal, P. Gangamohan, and B. Yegnanarayana, "Relative importance of different components of speech contributing to perception of emotion," in *Sixth International Conference on Speech Prosody*, 2012.

[3] M. Portnoff, "Time-scale modification of speech based on short-time fourier analysis," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 29, no. 3, pp. 374–390, 1981.

[4] N. Adiga, D. Govind, and S. M. Prasanna, "Significance of epoch identification accuracy for prosody modification," in *Signal Processing and Communications (SPCOM), 2014 International Conference on*. IEEE, 2014, pp. 1–6.

[5] K. S. Rao and B. Yegnanarayana, "Prosody modification using instants of significant excitation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 3, pp. 972–980, 2006.

[6] S. Prasanna, D. Govind, K. S. Rao, and B. Yenanarayana, "Fast prosody modification using instants of significant excitation," in *Proc Speech Prosody*, 2010.

[7] M. R. Thomas, J. Gudnason, and P. A. Naylor, "Application of dypsa algorithm to segmented time scale modification of speech," in *proc EUSIPCO*. IEEE, 2008.

[8] E. Moulines and F. Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," *Speech communication*, vol. 9, no. 5-6, pp. 453–467, 1990.

[9] P. Gangamohan, S. R. Kadiri, and B. Yegnanarayana, "Analysis of emotional speech at subsegmental level." in *INTERSPEECH*, 2013, pp. 1916–1920.

[10] H. K. Vydana, S. R. Kadiri, and A. K. Vuppala, "Vowel-based non-uniform prosody modification for emotion conversion," *Circuits, Systems, and Signal Processing*, vol. 35, no. 5, pp. 1643–1663, 2016.

[11] J. Tao, Y. Kang, and A. Li, "Prosody conversion from neutral speech to emotional speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1145–1154, 2006.

[12] D. Govind, S. M. Prasanna, and B. Yegnanarayana, "Neutral to target emotion conversion using source and suprasegmental information." in *Interspeech*, 2011, pp. 2969–2972.

[13] H. K. Vydana, V. Vidyadhara Raju, V, S. V. Gangashetty, and A. K. Vuppala, "Significance of emotionally significant regions of speech for emotive to neutral conversion," in *International Conference on Mining Intelligence and Knowledge Exploration*. Springer, 2015, pp. 287–296.

[14] V. Vidyadhara Raju, V, P. Gangamohan, S. V. Gangashetty, and A. K. Vuppala, "Application of prosody modification for speech recognition in different emotion conditions," in *TENCON*, 2016.

[15] S. G. Koolagudi, S. Maity, V. A. Kumar, S. Chakrabarti, and K. S. Rao, "IITKGP-SESC: Speech Database for Emotion Analysis," in *Contemporary Computing*. Springer, 2009, pp. 485–492.

[16] P. Gangamohan, V. K. Mittal, and B. Yegnanarayana, "A flexible analysis synthesis tool (fast) for studying the characteristic features of emotion in speech," in *Consumer Communications and Networking Conference (CCNC)*. IEEE, 2012, pp. 250–254.

[17] K. S. R. Murty and B. Yegnanarayana, "Epoch extraction from speech signals," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 8, pp. 1602–1613, 2008.

[18] A. K. Vuppala, J. Limmayya, and G. Raghavendra, "Neutral speech to anger speech conversion using prosody modification," in *Mining Intelligence and Knowledge Exploration*. Springer LNAI, 2013, pp. 383–390.

[19] N. Dhananjaya and B. Yegnanarayana, "Voiced/nonvoiced detection based on robustness of voiced epochs," *IEEE Signal Processing Letters*, vol. 17, no. 3, pp. 273–276, 2010.