Mobile Networks and Applications (2019) 24:193–201
https://doi.org/10.1007/s11036-018-1052-9

Application of Emotion Recognition and Modification for Emotional Telugu Speech Recognition

Vishnu Vidyadhara Raju Vegesna[1] · Krishna Gurugubelli[1] · Anil Kumar Vuppala[1]

Published online: 1 May 2018

Abstract

Majority of the automatic speech recognition systems (ASR) are trained with neutral speech and the performance of these systems are affected due to the presence of emotional content in the speech. The recognition of these emotions in human speech is considered to be the crucial aspect of human-machine interaction. The combined spectral and differenced prosody features are considered for the task of the emotion recognition in the first stage. The task of emotion recognition does not serve the sole purpose of improvement in the performance of an ASR system. Based on the recognized emotions from the input speech, the corresponding adapted emotive ASR model is selected for the evaluation in the second stage. This adapted emotive ASR model is built using the existing neutral and synthetically generated emotive speech using prosody modification method. In this work, the importance of emotion recognition block at the front-end along with the emotive speech adaptation to the ASR system models were studied. The speech samples from IIIT-H Telugu speech corpus were considered for building the large vocabulary ASR systems. The emotional speech samples from IITKGP-SESC Telugu corpus were used for the evaluation. The adapted emotive speech models have yielded better performance over the existing neutral speech models.

Keywords ASR · Emotion recognition · Emotive speech

1 Introduction

Human speech is considered to be the natural form of communication in the field of human-machine interactions. The human speech provides the various additional information about the speaker such as gender, accent, health, age, emotional state etc apart from the linguistic content. ASR system performance can be boosted up by considering this additional information provided by the speech signal. A mismatch in the training and testing (evaluating) conditions is observed in the majority of the real-life applications which affects the ASR system performance. The physical state of the speaker, context, dialect, speaking style and vocal effort are the intrinsic (speaker related) properties of the human speech responsible for the mismatch. The other non-speaker related (extrinsic) properties such as channel variations, background noise, recording environments etc also contribute to the mismatch in the training and testing conditions.

The emotional state of the speaker is one of the speaker related issues responsible for the degradation of ASR system performance. The emotions in human speech convey the paralinguistic information which reflects the changes in the physical and mental state of the speaker. These emotional states of the speaker provide the valuable feedback information in the human communication. These emotions contribute to the changes in the physiological aspects of respiration and the articulation of speech. These physiological changes are manifested in the prosody parameters such as pitch, duration and energy [1, 2]. In the recent years, there has been a growing interest towards the speech interfaces which handle the emotions in the user's voice [3]. So, recognizing the emotions from the human speech is considered to be the emerging research area in several speech interfaces. The performance of the

✉ Vishnu Vidyadhara Raju Vegesna
vishnu.raju@research.iiit.ac.in

Krishna Gurugubelli
krishna.gurugubelli@research.iiit.ac.in

Anil Kumar Vuppala
anil.vuppala@iiit.ac.in

[1] Speech Processing Lab, KCIS, International Institute of Information Technology, Hyderabad (IIIT-H), Hyderabad, India

emotion recognition systems is purely dependent on the differences between the training and testing scenarios. The classification performance is affected due to the speaker variations, dialects and small differences in the background noise. Hence the design of generalized and robust emotion classifiers are the challenges in the field of affective computing [4]. Different approaches were proposed to improve the robustness of the emotion classifiers namely: feature selection approach [5], the collection of natural databases [6, 7] and speaker normalization approach [8–10]. The model level adaptation [11, 12] is an effective approach in which the classifiers are modified in such a way that the gap between the training and testing conditions is reduced. A supervised domain adaptation approach [13] was proposed to address the classification problem by using a multi-corpus framework.

The major focus of the speech researchers was towards the recognition of emotions being expressed from the speech [14–16]. It solves only half of the problem in the human-machine interaction. The other half is about the ASR problem i.e. recognizing the verbal content of the spoken human speech [17]. The emotion factors of pitch, stress and pauses present in the human speech have a strong impact on the recognition performance [18]. The presence of different emotions of happiness, anger, frustration, impatience and grief in human speech are always misleading. The significant changes in the speech parameters are responsible for the degradation of the ASR performance. Exploring different techniques to improve the performance of ASR has become a hot research topic [19, 20].

There were different methods proposed in the literature to reduce the speaker dependency. The first method proposed was by estimating the warping factors of the utterances by performing the feature normalization. These methods were referred as vocal tract normalization (VTLN) methods [21, 22]. The next method proposed was adapting the acoustic models to the features of each utterance. This method was referred as maximum likelihood linear regression (MLLR) which has been used in the recent ASR systems [23]. The third method focuses on computing the features that are independent of these speaker characteristics.

In this work, the improvement of the degraded ASR system performance is shown at two stages. In the first stage emotion recognition is performed to recognize the emotion from the human speech. Differenced prosody features are considered for the task of emotion recognition. The recognized emotions from the human speech are passed on to the next stage. The adapted emotive ASR system is selected based on the detected emotions from the first stage. The ASR adaptation to the specific emotions is carried out by including the corresponding emotive speech along with the existing neutral speech. This emotive speech is generated from the existing neutral speech using the emotion conversion method proposed in [24, 25]. This emotion conversion is done by performing the prosody modification on the neutral speech which involves the process of altering the pitch contour and durations of the sound units without introducing the spectral and temporal distortions in human speech [26]. The influence of different emotions on various prosody parameters have been studied to perform prosody modification [27–30].

An analysis study is done in [24, 25] to capture the relative changes in these prosody parameters at much finer levels. These relative changes in the prosody components are reported as prosody modification factors. The emotive version of the input neutral speech is generated by considering the prosody modification factors at non-uniform level. The non-uniformity is addressed by considering the position of word segments occurring in a sentence (i.e. starting, middle and end segments). This work focuses on performing the automatic prosody modification on the detected emotions. The generated emotive version speech is used at the training phase of the ASR systems. An effort was put towards the collection of speech database for building a large vocabulary speech recognition system for the Telugu language which was named as IIIT-H Telugu corpus [31]. This Telugu corpus is used in our study for the purpose of building a neutral speech trained ASR system and IITKGP-SESC emotional speech corpus [32] is used for testing the ASR system.

The details of the database and the ASR system overview is explained in detail in Section 2. The proposed approach for the improvement in the performance of ASR systems is discussed in Section 3. The results obtained from this approach is discussed in detail in Section 4. Section 5 concludes the paper with information on the future scope of presented work.

2 Experimental Setup for ASR system

This section provides the details of the Telugu speech corpus and the baseline performance.

2.1 Telugu speech corpus

In this work, two databases were considered for the training and evaluation of ASR systems. IIIT-H Telugu speech corpus is used for the training of the ASR system and IITKGP-SESC is used for the evaluation purpose. This IIIT-H Telugu neutral speech corpus and the text corpus design have resulted from the joint efforts of Speech Processing Lab at IIIT-Hyderabad, India. IIIT-H Telugu speech corpus consists of data from 92 speakers and 64,464 utterances in total. These speech samples were collected in a noise-free recording studio with the help

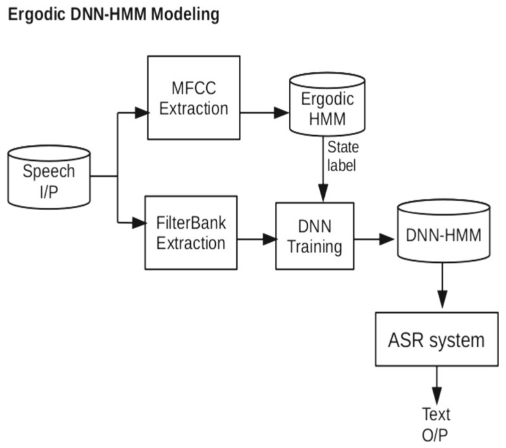

Fig. 1 System overview

of Zoom recorder. The duration of these speech samples is 10 hours with a sampling frequency of 16 kHz. This large vocabulary database consists of 25,700 unique Telugu words. IITKGP-SESC consists of 1500 utterances for the evaluation purpose. The recordings comprises of data 5 male and 5 female professional speakers. Each speaker has spoken out 15 sentences in eight different emotions. The basic emotion samples of anger, happy and compassion along with neutral speech are considered for the evaluation.

2.2 ASR system overview

The ergodic HMMs are trained along side using the Mel-frequency cepstral coefficients(MFCC) features and the classification is done using the maximum likelihood classification [33]. The forced alignment is performed for the generation of state labels from the information provided from the HMMs [34]. The DNN-HMM based experiments are implemented using Kaldi toolkit [35]. The overview of the ASR system is clearly shown in Fig. 1. This ASR system includes DNN-HMM acoustic modeling. To perform this model training, the processing of speech samples is done to extract the MFCC and log Mel-filterbank (FilterBank) features. For directly modeling the DNN should be trained with input FilterBank features and the HMM states are the learning target.

51 unique phones are considered for building the large vocabulary Telugu speech corpus. These phones are mapped on the 25,700 words to generate the word level models for this large vocabulary telugu ASR. The training of the neutral speech is done with the speech samples of 92 speakers for building the word level HMM-GMM acoustic model. The language model employed is a ARPA format trigram model which is built from SRILM tool kit. The testing is performed on emotional speech samples collected from 10 speakers of IITKGP-SESC corpus. The ASR is evaluated on different emotions of anger, happy and compassion.

The performance of the baseline ASR system is reported in terms of word error rate (WER) in Table 1. The results were reported for the 5-fold validation case. The combinational results for the triphone GMM-HMM acoustic models using linear discriminant analysis (LDA), maximum likelihood transform (MLLT) and speaker adaptive training were reported in Column 2 and 3 respectively. The results for the DNN and sub-space gaussian mixture models (SGMM) triphone acoustic models were reported in Column 4 and 5 respectively. It is observed that the WER is minimum for the case of acoustic model trained with SGMM which is shown in Column 5.

Table 1 has reported the performance of ASR system when evaluated on different emotions of neutral, anger, happiness and compassion. A degradation in the ASR system performance for the three emotions of anger, happiness and compassion. WER is minimum for the case of neutral emotion. The WER is less in the case of neutral speech as the ASR is trained on neutral speech. More degradation of ASR system performance is observed in the case of emotion of happiness speech.

3 Proposed approach for the improvement in the performance of ASR system

The proposed approach for the improvement in the performance of ASR system is shown in Fig. 2. Improvement

Table 1 WER of the ASR system trained on neutral speech and tested on different emotions of anger, happiness and compassion. Column 2 and 3 shows the WER for the triphone GMM-HMM acoustic model trained along with (LDA+MLLT) and (LDA+ MLLT+SAT) respectively. The WER for the triphone DNN-HMM model is shown in Column 4. Column 5 reports the WER for the SGMM trained models from the alignments generated from the HMM states

	Word Error Rate (%)			
Emotion	GMM-HMM Triphone (LDA+MLLT)	GMM-HMM Triphone (LDA+MLLT+SAT)	DNN-HMM	SGMM
Neutral	29.49	15.38	10.68	10.68
Anger	44.44	34.62	25.64	22.51
Happy	42.31	35.04	27.35	24.79
Compassion	32.48	26.92	21.37	18.67

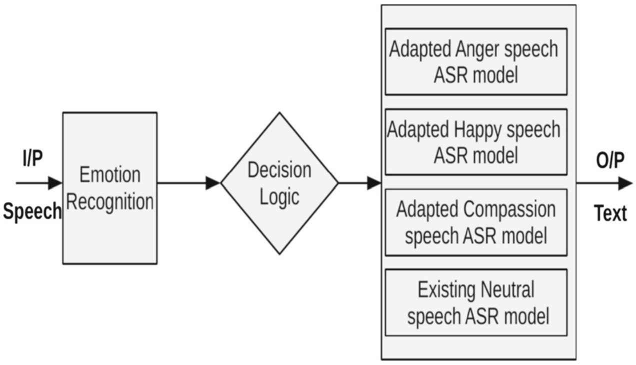
I/P
Speech
Emotion Recognition
Decision Logic
Adapted Anger speech ASR model
Adapted Happy speech ASR model
Adapted Compassion speech ASR model
Existing Neutral speech ASR model
O/P
Text

Fig. 2 Two stage approach for the improvement of ASR systems

in the ASR system is shown by implementing a two stage approach. In the first stage a emotion recognition block is introduced to detect the unknown emotions from the input speech. The information regarding the emotion recognition block is explained in detail in Section 3.1. These recognized emotions are passed to the next stage for the ASR model selection. The details of the emotive speech adaptation models are discussed in Section 3.2.

3.1 Emotion recognition

For the task of emotion recognition the combined features of vocal tract, prosody and differenced prosody features are considered. The combined MFCC and relative prosody features have yielded a 75% recognition rate for the given GMM-HMM classifier which has been discussed in the results Section 4. The reason for not considering the source features and the importance of excitation source, vocal tract and prosody features for the context of emotions is explained in detail in Sections 3.1.1 to 3.1.3.

3.1.1 Excitation source features

In the development of speech based emotion recognition systems the important issue is regarding the extraction of features that characterize the emotions. The classification performance is purely dependent on the selection of the suitable features of the human speech. Different speech features are responsible in representing the speaker and emotion information in highly overlapped manner. In emotional speech analysis the features are mostly selected on experimental basis. Hence the prosodic, vocal tract system and excitation source features are considered for the experimental study.

In order to extract the excitation source features, the information regarding the excitation source signal is obtained by suppressing vocal tract (VT) characteristics. The prediction of the vocal tract information is done by considering the linear prediction coefficients (LPCs) from the input speech. The vocal tract information is suppressed and separated by performing the inverse filter formulation. The resultant signal is referred as linear prediction residual [36] which contains the majority of excitation source information. The features derived from the LP residual are known as excitation source or sub-segmental or source features. The results obtained from the previous studies [37, 38] on excitation source features could not outperform the results of the well established prosodic and spectral features. In this work the prosodic and spectral features are considered for the task of emotion recognition in stage 1.

3.1.2 Vocal tract features

In general the vocal tract system features are extracted for a speech segment of length 20-30 ms. These vocal tract characteristics are seen in the frequency domain analysis of the input speech signal. The short time spectrum is obtained by applying the Fourier transform on the speech frame. This spectrum provides the information of the features like spectral energy, formants and slope. The Fourier transform on the log magnitude spectrum gives the cepstral domain of the speech frames [39]. The features extracted from this cepstral domain represent the vocal tract information. For the case of emotional speech, the sequence of shapes of vocal tract are responsible for producing the different sound units. MFCCs, linear prediction cepstral coefficients (LPCCs) are the well known spectral features used in the literature [14]. Generally, these spectral features are considered as the better correlates of the rate of change in the articulatory movements and the varying shapes of the vocal tract [40]. Block processing approach is used to extract the spectral features of the speech signal. The speech signal processed using frame by frame analysis where the frame size is 20 ms and the frame shift is 10 ms. In real-time, the emotional speech information is of more prominence in the syllabic regions (i.e consonants and vowels) or in the emotion salient regions (i.e words). This emotion specific speech information is purely dependent on the pattern of the emotion expressions. The changes due to the emotions are clearly observed in the much finer spectral variations. In this work MFCC spectral features are considered for the task of emotion recognition.

3.1.3 Prosody features

Duration, intonation and intensity patterns are the sequence of sound units imposed by human beings during the speech production. The natural communication through human speech is only possible through proper incorporation of the prosody constraints such as intonation, duration and intensity. Prosody deals with the larger units of speech such as syllables, words and sentences known as suprasegmental

Table 2 Emotion recognition results achieved on IITKGP-SESC dataset for the baseline feature sets on GMM-HMM and SVM classifier with “leave-one-speaker-out” cross validation with the combined proposed MFCC and differenced prosody features

Features	Excitation source	MFCC	Prosody	Differenced prosody	Combined MFCC and differenced prosody
Recognition (%)	52.37%	55.62%	62.22%	65.38%	75.28%

information. The patterns of the intonation(F_0), duration and energy are very useful in the acoustic representation of the prosody. These prosody components represent the speech perceptual properties [27]. All the previous studies have stated that the pitch, energy, duration and their derivatives are the components used as the higher acoustic correlates of emotions [41–43]. Recently differenced prosody features [44] have been explored for the task of emotion recognition which has shown better performance when compared to the conventional prosody features. The differenced prosody features are considered for the reason that the influence of the natural variations in different emotions is reduced in it. The procedure to extract the differenced prosody features is explained in detail in [44]. The prosody and differenced prosody features are considered for the task.

Speech samples from each speaker is collected in 10 different sessions. Four different emotions of anger, happiness, neutral and compassion are considered for the task of emotion recognition. A total of 6,000 utterances were considered for the experimentation. Results were reported on the leave one speaker out (LOSO) basis where a 10-fold cross validation is done.

39-dimensional MFCC features were used for the GMM-HMM classifier in this emotion recognition task. For the prosody features 11-dimensional feature vectors (i.e. maximum, mean and standard deviation for each of the pitch, strength of excitation (SoE) and energy along with average and duration ratio) are considered on the SVM classifier. The relative prosody features also have the same feature dimensions of the prosody features but the relative difference of the prosody features of the emotion speech with neutral speech is considered with the same SVM classifier. The emotion recognition results obtained using these MFCC, prosody, relative prosody features is shown in Table 2.

From the Table 2 it is evident that the combined MFCC and differenced prosody features have yielded better performance. For the GMM-HMM classifier, the 32 mixture model is used for a iteration count of around 75. In the case of SVM classifier, linear kernel is considered for the emotion recognition.

3.2 Emotive speech adaptation

After performing the task of emotion recognition at stage 1 the emotions in the speech samples are used for the selection of ASR models which are adapted to the specific emotions. During the adaptation, ASR models are trained with emotive speech obtained from neutral speech using prosody modification. The prosody modification involves the process of altering the pitch contour and durations of the sound units without introducing the spectral and temporal distortions in human speech [26].

Prosody components such as pitch, energy, duration and their derivatives are the higher acoustic correlates of emotions. So,the expressed human emotions can be captured through prosodic parameters. There are four different levels of manifestation that the prosody can be defined at linguistic, articulatory, acoustic and perceptual levels [45]. Prosody at linguistic level refers to the semantic emphasis on an element or relating the different linguistic elements. At articulatory level, it refers to series of articulatory movements which include the variations in their amplitudes and air pressure. The energy from the muscle movement from the respiratory lungs along with the vocal tract system is responsible for generation of the acoustic waves. Prosody at acoustic level means

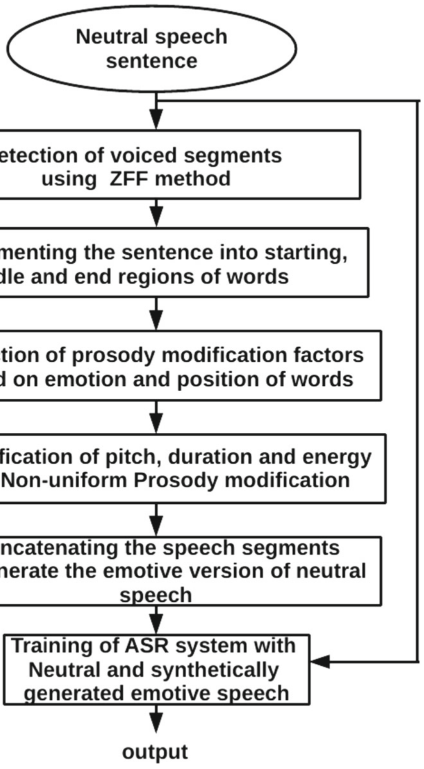
Neutral speech sentence
Detection of voiced segments using ZFF method
Segmenting the sentence into starting, middle and end regions of words
Selection of prosody modification factors based on emotion and position of words
Modification of pitch, duration and energy by Non-uniform Prosody modification
Concatenating the speech segments to generate the emotive version of neutral speech
Training of ASR system with Neutral and synthetically generated emotive speech
output

Fig. 3 Steps to perform prosody modification

Table 3 Non-uniform male prosody modification factors

Modification factors	Starting words of a sentence	Middle words of a sentence	Ending words of a sentence
Anger modification factors			
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

the analysis of acoustic components of fundamental frequency(F_0), duration, energy and intensity. At the perception level, prosody means the subjective evaluation or experience of the listener which include pauses, melody of the speech. The main reason for considering only the acoustic properties of the human speech is as it becomes difficult in analyzing prosody through speech production or perception mechanism. The procedure to perform the prosody modification is shown in Fig. 3.

The prosody modification approach proposed in [46] has been used and the steps involved to perform the emotion conversion is followed. In the first step from the neutral speech, segmentation of the entire utterance is done by detecting the voiced segments using ZFF based method [47, 48] from the epoch locations. The regions between the two consecutive pauses or silence is considered to be a acoustic word. Segmentation is done in such a way that there exists a minimum of one acoustic word in every starting, middle and end regions of the utterance. The utterances which do not have at least one acoustic word in any of the regions are neglected for the task of prosody modification.

A suprasegmental analysis is introduced to perform the emotion conversion. The neutral utterance is divided into three supra-segments as starting, middle and ending words. An analysis study is done on the neutral speech with respect to the emotive speech to estimate the prosodic changes of pitch and duration. The relative changes in these prosody parameters are studied at gross and finer levels. These relative changes in the prosody parameters of pitch, duration and energy are reported as prosody modification factors. middle and end regions of words. The pitch,duration and energy parameters of the emotional sentence is modified by the specific modification factors which are mentioned in Tables 3 and 4. The male prosody modification factors are reported in Table 3 and the female prosody modification factors in Table 4. The expression required to convert the

Table 4 Non-uniform prosody female modification factors

Modification factors	Starting words of a sentence	Middle words of a sentence	Ending words of a sentence
Anger modification factors			
Pitch	0.81	0.83	0.88
Duration	1.36	1.26	1.45
Energy	1.05	0.95	0.94
Happy modification factors			
Pitch	0.89	0.87	0.89
Duration	1.01	0.90	1.18
Energy	0.98	0.95	0.99
Compassion modification factors			
Pitch	1.15	1.06	0.98
Duration	0.91	1.04	1.33
Energy	1.03	1.04	1.06

Table 5 WER of the ASR system trained on neutral speech and tested on different emotions of anger, happiness and compassion. Column 2 shows the WER for the Baseline ASR system. The WER of ASR system when only the selection of adapted emotive ASR models is done based on the prior known information of the emotional speech is shown in Column 3. Column 4 reports the WER for the ASR system which has both emotion recognition and adapted emotive speech ASR models

	Word Error Rate (%)		
Emotion	Baseline without	Adapted Emotive ASR from prior emotion information	Proposed approach
Neutral	10.68	10.68	10.68
Anger	22.51	17.09	18.98
Happy	24.79	13.25	15.13
Compassion	18.67	15.17	16.08

prosody parameters of neutral speech to emotive speech is given as below,

$$Prosody_{(emotive)} = Prosody_{(neutral)} * ModificationFactor \quad (1)$$

These modification factor values for the corresponding male and female are picked and applied to generate the neutral version speech from the given emotional speech. All the prosody modified speech segments are concatenated and synthesized to form the neutral speech. The output prosody modified speech generated from the stage 2 is passed to the existing ASR system for testing purpose.

4 Results

In this section the results obtained at the stage 1 and stage 2 are discussed in detail. The emotion recognition results obtained at stage 1 is discussed in Table 2. The speech samples from IITKGP-SESC emotion corpus are considered for the evaluation.

4.1 Performance of ASR system on the proposed approach

The WER obtained on the ASR system after performing the emotion recognition and adapted ASR model selection at stage 1 and stage 2 is shown in Table 5. The baseline results shown in column 2 of Table 5 indicate the WER obtained on the ASR system without performing the emotion recognition task. The outline of this approach is shown in Fig. 4. The results shown in the Table 5 clearly indicate the degradation of the ASR system performance in the presence of emotions.

Column 3 of Table 5 indicates the the WER of the ASR system obtained when the selection of the adapted emotive ASR is done from the prior knowledge of the emotional speech. An improvement in the ASR performance is clearly visible in column 3 when compared to results in column 2. In real-time there is no provision in identifying the specific emotions from the content of the emotional speech directly. Hence a proper emotion recognition block is required to be implemented before the selection of adapted emotive ASR systems. Results obtained from Table 2 clearly shows an 75% recognition rate for all the emotions. The recognized emotions are passed to the next level in which the selection of the adapted ASR models is done automatically from the information of the emotion recognition block. The outline of this approach is shown in Fig. 5.

From Table 5, the best performance was observed for the case when the emotion information is priorly known and the prosody modification factors picked based on the prior information. In real-time emotional speech processing applications, the emotion recognition block alongside the selection of adapted Emotive ASR models have yielded better results. The best performance was observed for the case of happy emotion when compared to other emotions.

These experiments were carried out initially at the testing phase of the ASR systems in which no better improvement

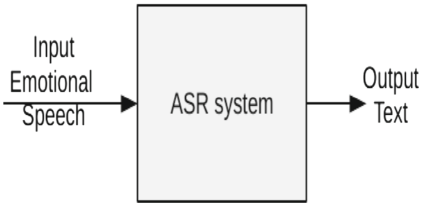

Fig. 4 ASR system in the absence of emotion recognition and emotive speech adapted ASR models

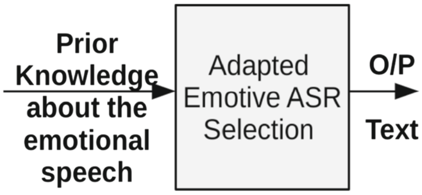

Fig. 5 ASR system when the specific emotive information is priorly known with the emotive speech adapted ASR models

was seen. The reason is that there is no proper control over the testing environments since there would be a lot of mismatch environments responsible for the degradation. Hence the proposed approach adaptation of these emotional speech models is done at the training phase in which the ASR systems are most robust to these emotional mismatch environments.

5 Conclusion

In this work the performance of Telugu ASR system is investigated in different emotive conditions. A significant degradation in the ASR system performance was observed in the presence of emotional speech. The performance was improved by adapting the emotional speech models to the existing ASR system. In practice, the prior information regarding the emotion specific state in the human speech is unknown to the select the required emotive specific ASR models. This problem is addressed by incorporating an emotion recognition block as a front-end to the ASR system. Based on the emotions recognized from the front-end, the selection of the corresponding adapted emotive ASR model was done. The front-end emotion recognition block was implemented with the combined differenced prosodic and vocal tract features. The importance of emotion recognition block at the front-end and adaptation of ASR system towards different emotions was observed in this work. Better feature representation for emotion recognition and different emotion adaptation strategies can be explored in the future work to improve the robustness of ASR system.

References

1. Gangamohan P, Mittal V, Yegnanarayana B (2012) Relative importance of different components of speech contributing to perception of emotion. In: Proc of Sixth international conference on speech prosody, China
2. YeonWoo Lee MK, Cheeyong K (2017) A study on colors and emotions of video contents-focusing on depression scale through analysis of commercials. Journal of Multimedia Information Systems 4(4):301–306
3. Dybkjaer L, Bernsen NO, Minker W (2004) Evaluation and usability of multimodal spoken language dialogue systems. Speech Comm 43(1-2):33–54
4. Busso C, Bulut M, Narayanan S, Gratch J, Marsella S (2013) Toward effective automatic recognition systems of emotion in speech. In: Social emotions in nature and artifact: emotions in human and human-computer interaction, pp 110–127
5. Busso C, Lee S, Narayanan S (2009) Analysis of emotionally salient aspects of fundamental frequency for emotion detection. IEEE transactions on audio, speech, and language processing 17(4):582–596
6. McKeown G, Valstar M, Cowie R, Pantic M, Schroder M (2012) The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. IEEE Trans Affect Comput 3(1):5–17
7. Mariooryad S, Lotfian R, Busso C (2014) Building a naturalistic emotional speech corpus by retrieving expressive behaviors from existing speech corpora. In: Proc of Fifteenth annual conference of the international speech communication association
8. Schuller B, Vlasenko B, Eyben F, Wollmer M, Stuhlsatz A, Wendemuth A, Rigoll G (2010) Cross-corpus acoustic emotion recognition: variances and strategies. IEEE Trans Affect Comput 1(2):119–131
9. Sethu V, Ambikairajah E, Epps J (2007) Speaker normalisation for speech-based emotion detection. In: Proc of 15th international conference on digital signal processing, pp 611–614
10. Busso C, Metallinou A, Narayanan SS (2011) Iterative feature normalization for emotional speech detection. In: Proc of international conference on acoustics, speech and signal processing (ICASSP), IEEE, pp 5692–5695
11. Deng J, Zhang Z, Marchi E, Schuller B (2013) Sparse autoencoder-based feature transfer learning for speech emotion recognition. In: Proc of humaine association conference on affective computing and intelligent interaction (ACII), IEEE, pp 511–516
12. Maeireizo B, Litman D, Hwa R (2004) Co-training for predicting emotions with spoken dialogue data. In: Proc of the interactive poster and demonstration sessions, ACL, p 28
13. Abdelwahab M, Busso C (2015) Supervised domain adaptation for emotion recognition from speech. In: Proc of International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, pp 5058–5062
14. Ververidis D, Kotropoulos C (2006) Emotional speech recognition: resources, features, and methods. Speech Comm 48(9):1162–1181
15. Schuller B, Seppi D, Batliner A, Maier A, Steidl S (2007) Towards more reality in the recognition of emotional speech. In: Proc of international conference on acoustics, speech and signal processing, vol 4. IEEE, pp IV–941
16. Schuller B, Batliner A, Steidl S, Seppi D (2009) Emotion recognition from speech: putting asr in the loop. In: Proc of international conference on acoustics, speech and signal processing (ICASSP), IEEE, pp 4585–4588
17. Athanaselis T, Bakamidis S, Dologlou I, Cowie R, Douglas-Cowie E, Cox C (2005) Asr for emotional speech: clarifying the issues and enhancing performance. Neural Netw 18(4):437–444
18. Steeneken HJ, Hansen JH (1999) Speech under stress conditions: overview of the effect on speech production and on system performance. In: Proc of international conference on acoustics, speech, and signal processing(ICASSP), vol 4. IEEE, pp 2079–2082
19. Benzeghiba M, De Mori R, Deroo O, Dupont S, Erbes T, Jouvet D, Fissore L, Laface P, Mertins A, Ris C et al (2007) Automatic speech recognition and speech variability: a review. Speech Comm 49(10-11):763–786
20. Sheikhan M, Gharavian D, Ashoftedel F (2012) Using dtw neural-based mfcc warping to improve emotional speech recognition. Springer journal on Neural Computing and Applications 21(7):1765–1773
21. Welling L, Ney H, Kanthak S (2002) Speaker adaptive modeling by vocal tract normalization. IEEE Transactions on Speech and Audio Processing 10(6):415–426
22. Sinha R, Umesh S (2002) Non-uniform scaling based speaker normalization. In: Proc of international conference on acoustics, speech, and signal processing (ICASSP), vol 1. IEEE, pp I–589
23. Müller F, Mertins A (2011) Contextual invariant-integration features for improved speaker-independent speech recognition. Speech Comm 53(6):830–841

24. Vydana HK, Vidyadhara Raju V, Gangashetty SV, Vuppala AK (2015) Significance of emotionally significant regions of speech for emotive to neutral conversion. In: Proc of international conference on mining intelligence and knowledge exploration, Springer, Hyderabad, pp 287–296
25. Vidyadhara Raju V, Vydana Vhk, Gangashetty SV, Vuppala AK (2017) Importance of non-uniform prosody modification for speech recognition in emotion conditions. In: Proc of Asia-Pacific Signal and information processing association annual summit and conference (APSIPA), IEEE
26. Adiga N, Govind D, Prasanna SM (2014) Significance of epoch identification accuracy for prosody modification. In: Proc of SPCOM, IEEE, Bangalore, pp 1–6
27. Rao KS, Yegnanarayana B (2006) Prosody modification using instants of significant excitation. IEEE Trans Audio Speech Lang Process 14(3):972–980
28. Tao J, Kang Y, Li A (2006) Prosody conversion from neutral speech to emotional speech. IEEE Trans Audio Speech Lang Process 14(4):1145–1154
29. Prasanna S, Govind D, Rao KS, Yenanarayana B (2010) Fast prosody modification using instants of significant excitation. In: Proc of speech prosody, Chicago
30. Thomas MR, Gudnason J, Naylor PA (2008) Application of dypsa algorithm to segmented time scale modification of speech. In: Proc of EUSIPCO, IEEE, Switzerland
31. Vidyadhara Raju VV, Gurugubelli K, Vydana HK, Pulugandla B, Shrivastava M, Vuppala AK (2017) Dnn-hmm acoustic modeling for large vocabulary telugu speech recognition. In: Proc of international conference on mining intelligence and knowledge exploration, Springer, pp 189–197
32. Koolagudi SG, Maity S, Kumar VA, Chakrabarti S, Rao KS (2009) IITKGP-SESC: speech database for emotion analysis. In: Contemporary computing, Springer, pp 485–492
33. Saul LK, Rahim MG (2000) Maximum likelihood and minimum classification error factor analysis for automatic speech recognition. IEEE Transactions on Speech and Audio Processing 8(2):115–125
34. Young S, Evermann G, Gales M, Hain T, Kershaw D, Liu X, Moore G, Odell J, Ollason D, Povey D et al (2002) The htk book. Cambridge university engineering department 3:175
35. Povey D, Ghoshal A, Boulianne G, Burget L, Glembek O, Goel N, Hannemann M, Motlicek P, Qian Y, Schwarz P et al (2011) The kaldi speech recognition toolkit. In: IEEE 2011 workshop on automatic speech recognition and understanding, IEEE Signal Processing Society
36. Makhoul J (1975) Linear prediction: a tutorial review. Proc IEEE 63(4):561–580
37. Koolagudi SG, Rao KS (2012) Emotion recognition from speech: a review. Int J Speech Technol 15(2):99–117
38. El Ayadi M, Kamel MS, Karray F (2011) Survey on speech emotion recognition: features, classification schemes, and databases. Pattern Recogn 44(3):572–587
39. Rabiner LR, Juang B-H (1993) Fundamentals of speech recognition. PTR Prentice Hall Englewood Cliffs, vol 14
40. Benesty J, Chen J, Huang Y (2008) Microphone array signal processing. Springer Science & Business Media, vol 1
41. Cowie R, Cornelius RR (2003) Describing the emotional states that are expressed in speech. Speech Comm 40(1-2):5–32
42. Bänziger T, Scherer KR (2005) The role of intonation in emotional expressions. Speech Comm 46(3-4):252–267
43. Lee CM, Narayanan SS (2005) Toward detecting emotions in spoken dialogs. IEEE Transactions on Speech and Audio Processing 13(2):293–303
44. Vidyadhara Raju VV, Gurugubelli K, Vuppala AK (2018) Differenced prosody features from normal and stressed regions foremotion recognition. In: 5th international conference on signal processing and integrated networks (SPIN), IEEE
45. Werner S, Keller E (1995) Prosodic aspects of speech. In: Fundamentals of speech synthesis and speech recognition, Wiley Ltd., pp 23–40
46. Govind D, Prasanna SM (2018) Prosody modification for speech recognition in emotionally mismatched conditions. Int J Speech Technol
47. Murty KSR, Yegnanarayana B (2008) Epoch extraction from speech signals. IEEE Trans Audio Speech Lang Process 16(8):1602–1613
48. Dhananjaya N, Yegnanarayana B (2010) Voiced/nonvoiced detection based on robustness of voiced epochs. IEEE Signal Process Lett 17(3):273–276