Speech Recognition under Stress Condition

Sumitra Shukla, S. R. Mahadeva Prasanna and S. Dandapat
Department of Electronics and Communication Engineering
Indian Institute of Technology Guwahati
Guwahati-781039, Assam, India
{sumitra, prasanna, samaren}@iitg.ernet.in

***Abstract*—The objective of this work is to conduct a speech recognition study and evaluate the performance of the same under stressed condition. The speech recognition study is conducted both in isolated word recognition and keyword spotting approaches. The word models are built during training using speech collected from neutral condition. During testing these models are tested with speech signals collected under different stressed conditions to quantify the amount of degradation in each stress condition. It is observed that the performance of the speech recognition system decreases significantly under stressed condition.**

I. Introduction

Speech is a complex signal which encodes message as well as paralinguistic information like speaker, emotion, acoustic environment, person's intention, language, accent and dialect [1]. Stress refers to the psychological state of the person due to internally induced factors like emotions or externally induced factors like Lombard effect. In human-human interaction, listener can recognize message as well as paralinguistic aspects present in the speech. At the same time the listener can also effortlessly extract only wanted information from the speech and neglect the rest by what is called selective attention. This is not understood well to mimic the same in human-computer interaction. Hence in the case of human-computer interaction, the performance of the system degrades because of the inability of the system to deemphasize the paralinguistic information. For instance, under stress the speech production varies with respect to neutral condition due to the constriction of various muscle structures present in the speech production system. This leads to the change in the characteristic of speech signal compared to the neutral condition. Identification of stress and properly compensating the same will give significant improvement in the performance of speech or speaker recognition systems. For this it is better to quantify the amount of degradation that will be caused due to the stressed condition. The present work deals with the quantification of degradation in the performance of speech recognition system under stressed condition.

Most of the earlier attempts in the stressed speech processing area focused on how to classify and compensate different stress conditions. To find the effect of stress in speech, researchers have studied the effects of stress at sentence, word and sound unit levels [1]. In these study they have analyzed percentage deviation of duration, intensity, glottal pulse shaping and vocal tract spectrum [1]. In some of the studies, speech recognizer was trained with neutral speech and during testing the effect of stress was compensated [2]. Under this condition, compensation techniques used for such analysis are formant location and bandwidth stress equalization [3], [4], [5], whole word cepstral compensation [6], slope-dependent weighting [7], formant shifting [8], source-generator based codebook stress compensation [9], [10], source-generator based adaptive cepstral compensation [11], [10]. The purpose of these studies are to improve the performance of speech recognition system. All these studies are based on the fact that under stressed condition the performance of the speech recognition system degrades, but not exactly to quantify how much degradation takes place. Even though it is a known fact that under stressed condition performance of the system degrades, it may be better to have first hand quantification of amount of degradation. Such quantification will help in the following ways: We will understand the exact amount of degradation caused by each stress condition. Accordingly methods may be developed to compensate the stress for each condition. This is the motivation for the present work.

In this study, we quantify the effect of stress in an automatic speech recognition task realized in isolated word recognition (IWR) and keyword spotting (KWS) tasks. The main intention behind using IWR and KWS approaches is that they are simple to implement. Further, the performance obtained can be easily attributed to the stress conditions. IWR deals with recognition of isolated words present in the test speech. Keyword spotting deals with recognition of keywords present in the continuous speech. Further the main focus of this paper is to understand the influence of stress in the isolated words than in continuous sentences. For feature extraction, mel-frequency cepstral coefficients (MFCC) are used and for modeling of speech vector quantization (VQ) technique is used. The remaining paper is organized as follows: In Section II, we describe the database used for the present work, brief introduction of isolated word recognition and keyword spotting in continuous speech with experimental details. Results and discussion of IWR system and KWS in continuous speech are described in Section III . Finally, summary and conclusions of the present work and also scope for the future work are mentioned in Section IV.

II. Speech Recognition under Stressed Condition

A. *Database*

The database employed for this study is SUSAS (speech under simulated and actual stress) database having isolated words

[12] and the database collected from our laboratory named as SUSE (speech under simulated emotion) for continuous speech [13]. The emotions considered in the SUSAS database are neutral, angry, loud, fast, slow, Lombard and soft.In SUSE database the emotions considered are neutral, happy, angry, loud, compassion, sad, bore and fear. Words used for isolated word recognition are break, change, degree, destination, east and eight. In keyword spotting task *India* and *won the* present in the continuous speech for the sentence *India won the match* are taken as keyword. All speech samples are sampled at 8 kHz sampling frequency with 16 bits/ sample resolution.

B. *Isolated Word Recognition*

Isolated word recognition deals with recognition of words which are uttered in isolated manner in both during training and testing. In this study, we quantify the effect of stress in isolated words. Because of stress, speech signal characteristics will deviate from the neutral condition. In order to find out the amount of deviation, neutral words are used for training the recognizer. IWR study is conducted for two cases. In the first case, isolated words are taken from SUSAS database for training and testing. In the second case, manually segmented words from the continuous speech in SUSE database are used for training and testing. Another important factor is the testing words are from neutral as well as different emotions.

IWR system is developed using MFCC algorithm for feature extraction where size of frame is 160 samples (i.e. 20 ms at 8 kHz) with frame shift of 80 samples (i.e. 10 ms at 8 kHz). The dimension of feature vectors is 13, excluding c_0. A model for each word is developed using VQ technique from the speech collected under neutral condition. The codebook of size 256 is considered in this study. This is because the total number of frames for each model is in the range of 4000-6000. During testing, test input is taken from neutral and from other emotions and compared with codebook of each word. The decision criteria used is minimum Euclidean distance to compare the feature vectors of test speech data to the codebook of the isolated word. Input word is considered as a hit, if maximum frames of that input word give minimum distance with that particular isolated word, when compared with codebook of other words.

C. *Keyword Spotting in Continuous Speech*

In continuous speech recognition the system recognizes an unconstrained speech, that is, speech where there is no pause between words. Alternatively, the role of keyword spotting system is to identify only meaningful words in the sentence and ignore the remaining words in the sentence. Since these words are not recorded in isolated manner, the characteristics of these words may be changed due to coarticulation of previous words. When the speakers are under stress, they will not give equal emphasis to each word, that is, percentage deviation of each of the words is not same in the whole sentence. Therefore by conducting KWS study we quantify how the stress will affect the performance of the system.

In this study, the continuous speech for the sentence *India won the match* is selected from the SUSE database and in this sentence *India* and *Won the* are taken as keywords and *Match* as a filler model. The training speech is manually segmented as mentioned earlier. The input speech is then parameterized as 13 MFCC coefficients per frame. Each frame is taken for duration of 20 ms with an overlap of 10 ms using neutral emotion. KWS system is developed using VQ technique where size of codebook of each keyword and filler is chosen as 128. During testing phase, as explained earlier, minimum distances with respect to all codebooks are obtained. In the test input, it is observed that some non-speech region also give minimum distance with some codebook which will increase the false alarm rate. In order to avoid this, speech/non-speech detection technique is employed by considering average of frame energies as a threshold. If energy value of the frame is less than threshold value, then that frame is considered as non-speech, otherwise speech frame. We assign index 1 for distance with respect to codebook of keyword *India*. Similarly, index 2 for keyword *won the* and index 3 for filler *match*. Considering frame of length of 160 samples of these three distances, depending on which of the three gives minimum distance, the corresponding index is assigned for that frame. Then finally distance curve is modified in the form of index either 1 or 2 or 3. Now a threshold equal to average duration of each keyword under neutral condition has to be taken. If the duration of the keyword assigned segment is greater than the threshold, then we can say that a keyword has occurred in that segment.

III. Results and Discussion

The performance of the speech recognition system may be measured in terms of detection rate. Detection Rate (DR) can be defined as the ratio of correctly detected words (M) to the total number of words occurred during testing (N).

Table 1 gives the performance of IWR system using SUSAS database. When we consider the average performance under all emotions, excluding neutral emotion, the performance of the system degrades for all the words. Further an average on the individual averages is 91.13% taken across all words and emotions, excluding neutral condition 99%. This result infers that there is significant degradation in the performance of IWR system under stressed condition. It is also observed that some stressed conditions like Angry, Loud and Soft the average performance is more degraded as compare to Clear, Fast Cond50 and Cond70. This infers that, effect of the stress on speech signal is not uniform under all stressed condition. It can also observed that under one stressed condition like angry the percentage degradation across all words are also not same which indicate that under stressed condition, speakers give non uniform emphasis to phonemes.

To confirm about this trend, one more IWR experiment was conducted by selected isolated words from SUSE database. In this database the speech under different stress conditions

TABLE I
PERFORMANCE OF IWR SYSTEM USING SUSAS DATABASE. ALL THE PERFORMANCES ARE IN (%)

Emotion	Break	Change	Degree	Destination	East	Eight	Avg. of each emotion
Neutral	100	100	100	100	100	94.44	99
Angry	88.89	94.44	38.89	83.33	55.56	66.67	71.29
Clear	94.44	100	94.44	100	100	83.33	95.37
Cond50	100	100	100	100	100	94.44	99
Cond70	100	100	100	100	83.33	83.33	94
Fast	100	100	94.4	100	100	100	99
Lombard	100	100	61.11	100	88.89	100	92
Loud	100	100	61.11	94.44	44.44	66.67	78
Question	100	94.94	100	100	94.94	100	98
Slow	100	100	94.44	100	100	72.22	94
Soft	88.89	94.44	100	100	100	55.56	89.82
Avg. perf. of all emotions	97.22	98.38	84.44	97.78	86.72	82.22	91.13

TABLE II
PERFORMANCE OF IWR USING SUSE DATABASE. ALL PERFORMANCES ARE IN (%)

Emotion	India	Won The	Avg. Perf. of each emotion
Neutral	100	100	100
Angry	81.82	72.73	77.27
Bore	90.91	90.91	90.91
Compassion	100	90.91	95.45
Fear	90.91	81.82	86.36
Happy	100	81.82	90.91
Loud	81.82	90.91	86.36
Sad	90.91	90.91	90.91
Avg. perf. of all emotions	90.91	85.72	88.31

TABLE III
PERFORMANCE OF KWS IN CONTINUOUS SPEECH USING SUSE DATABASE. ALL THE PERFORMANCES ARE IN (%)

Emotion	India	Won The	Avg. Perf. of each emotion
Neutral	90.91	100	95.455
Angry	45.45	72.73	59.09
Bore	72.73	100	86.365
Compassion	81.82	72.73	77.275
Fear	72.73	54.55	63.64
Happy	81.82	72.73	77.275
Loud	36.36	72.73	54.545
Sad	90.91	90.91	90.91
Avg. perf. of all emotions	68.83	76.63	72.73

is recorded as continuous speech. For comparison purpose, the words are manually segmented, both for training and testing.The performance of the IWR system is given in Table 2. In this case also the same trend continues. The overall average performance is 88.31% compared to 100% under neutral condition. The average performance across each emotion also shows wide variation compared to the average value of 88.31%. For instance the same angry emotion gives 77.27%, which is the worst performance across different stress conditions as in the earlier case. Thus this study reconfirms the all trend observed in the earlier experiment. After comparing table 1 and table 2, it can be found that SUSE database performance is more degraded than SUSAS database. Since these words are manually segmented from the continuous speech, hence due to coarticulation performance is further reduced.

In Table 3 we can see that the performance of KWS in continuous speech is degraded even more compared to the performance of IWR for the same data. This is mainly because in this system the word is assigned to be a keyword if consecutive frames cross the threshold value. Here threshold criteria for detection are average duration of keyword obtained from neutral speech. Since under stress condition, the average duration of keyword reduces and therefore some of the keywords are not properly detected. Thus the stress condition not only affects the individual word level performance, but also the degradation in various supra-segmental features like duration, energy and so on.

IV. SUMMARY AND CONCLUSIONS

In this work we have quantitatively demonstrated the effect of stress in the speech recognition task. The IWR and KWS approaches are selected for the demonstrating. The word models are built during training using speech collected under neutral emotion. During testing these models are tested using speech collected under different emotions. To quantitatively demonstrate the amount of degradation that occurs under different emotions. The objective of this work is to quantitatively evaluate the degradation in the performance for speech recognition under stressed condition. For this MFCC and VQ techniques are used. Since VQ technique does not give sequence information, it fails in confusable word modeling. However, in present study the words are distinct from each other and modeled with same technique therefore the degradation is only due to stress. From this study, it is observed that MFCC feature give large variation due to stress, hence different feature extraction method and modeling technique may be exploit to improve speech recognition performance. these are to be addressed in our future work.

REFERENCES

[1] John H. L. Hansen and Sanjay Patil, "Speech under Stress: Analysis, modeling and Recognition," Lecture notes in computer science, Springer Berlin, vol. 4343, pp.108-137, 2007.

[2] John H. L. Hansen and Sahar, "Robust Speech Recognition Training via Duration and Spectral -based Stress Token Generation," *IEEE Trans. Speech Audio Proc.*, vol. 3, pp. 415-421, Sep. 1995.

[3] John H. L. Hansen and M. A. Clements, "Stress compensation and noise reduction algorithms for robust speech recognition", in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Proc.*, Glasgow, U.K., May 1989,pp. 266-269.

[4] John H. L. Hansen and A. Clements, "Source Generator Equalization and Enhancement of Spectral Properties for Robust Speech Recognition in Noise and Stress," *IEEE Trans. Speech Audio Proc.*, vol. 3, pp. 407-415, Sept. 1995.

[5] John H. L. Hansen, "Analysis and Compensation of Speech under Stress and Noise for Environment Robustness in Speech Recognition," *Speech Comm.*, vol. 20, pp. 151-173, June 1996.

[6] Y. Chen, "Cepstral domain stress compensation for robust speech recognition," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Proc.*, Dallas, TX, Apr. 1987, pp. 717-720.

[7] B. Stanton, L. Jamieson, and G. Allen, "Robust recognition of loud and Lombard speech in the fighter cockpit environment," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Proc.*, Glasgow, U.K., May 1989,pp. 675-678.

[8] Y. Takizawa and M. Hamada, "Lombard speech recognition by formantfrequency- shifter LPC cepstrum," in *Proc. Int. Conf. Spoken Language Proc.*, Kobe, Japan, Nov. 1990, pp. 293-296.

[9] J. H. L. Hansen and O. N. Bria, "Lombard effect compensation for robust automatic speech recognition in noise," in *Proc. Int. Conf. Spoken Language Proc.*, Kobe, Japan, Nov. 1990, pp. 1125-1128.

[10] J. H. L. Hansen, "Analysis and compensation of speech under stress and noise for environmental robustness in speech recognition,," *Speech Comm.*, vol. 20, pp.151-173, Nov. 1996.

[11] J. H. L. Hansen, "Morphological constrained enhancement with adaptive cepstral compensation (MCE-ACC) for speech recognition in noise and Lombard effect," *IEEE Trans. Speech Audio Proc.*, vol. 2, pp. 598-614, Oct. 1994.

[12] J. H. L. Hansen and S. Bou-Ghazale, "Getting Started with SUSAS: A Speech Under Simulated and Actual Stress Database," in *Proc. EUROSPEECH-97*, Vol.4, pp. 1743-1746, Rhodes, Greece, Spetember 1997.

[13] S. Ramamohan and S. Dandapat, "Sinusoidal model-based analysis and classification of stressed speech," *IEEE Transactions on Audio, Speech and Language Proc.*, vol. 14, pp. 737 - 746, May 2006.