Stressed speech recognition using a warped frequency scale

D. Gharavian[1,2a)] **and S. M. Ahadi**[1]

[1] *Electrical Engineering Department, Amirkabir University of Technology,*
Hafez Avenue, Tehran 15914, Iran

[2] *Power Water University of Technology, Tehranpars Hakimieh,*
P.O. Box 16756–1719, Tehran, Iran

a) *gharavian@pwut.ac.ir*

Abstract: The use of emotion-initiated gestures in human speech communication results in the improvement of speech understanding. However, this is a source of difficulty for automatic speech recognizers. In this paper, using the orderly changes found in the second formant, due to stress, a warping function is introduced that can be applied to the mel frequency scale during the calculation of MFCC parameters. We show that this approach leads to improvements in the stressed speech recognition results. Furthermore, using the second formant frequency as an extra element of the feature vector leads to further improvements in the speech recognizer performance.

Keywords: prosody, stress, speech recognition, frequency warping, formants

Classification: Science and engineering for electronics

References

[1] M. Chu, Y. Wang, and L. He, “Labeling Stress in Continuous Mandarin Speech Perceptually,” *Proc. of the 15th International Congress of Phonetic Sciences*, Barcelona, 2003.

[2] D. Gharavian and S. M. Ahadi, “Statistical Evaluation of The Influence of Stress on Pitch Frequency And Phoneme Durations in Farsi Language,” *Proc. of the EUROSPEECH’03*.

[3] D. Gharavian and S. M. Ahadi, “Evaluation of The Effect of Stress on Formants in Farsi Vowels,” *Proc. of the ICASSP’04*, Montreal.

[4] D. Gharavian, “Prosody in Farsi Language and Its Use In Recognition of Intonation and Speech,” *PhD Thesis, Elec. Eng. Dept., Amirkabir University of Technology*, Tehran (in Farsi), 2004.

[5] S. J. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK Book (ver 3.2)*, Cambridge University Eng. Dept, 2002.

[6] E. Eide and H. Gish, “A Parametric Approach to Vocal Tract Length Normalization,” *Proc. of the ICASSP’96*, Atlanta, USA, pp. 346–348.

[7] M. Bijankhan, J. Sheikhzadegan, M. R., Roohani, Y. Samareh, C. Lucas, and M. Tebiani, “The Speech Database of Farsi Spoken Language,” *Proc. of the SST ’94*, Perth, Australia.

[8] S. S. McCandless, “An Algorithm for Formant Extraction Using Linear

DOI: 10.1587/elex.5.187
Received January 01, 2008
Accepted February 06, 2008
Published March 25, 2008

Prediction Spectra," *IEEE Trans. on Acoustics, Speech and Signal Processing, ASSP-22*, no. 2, pp. 135–141, 1974.

[9] D. Gharavian and S. M. Ahadi, "Use of Formants in Stressed and Unstressed Continuous Speech Recognition," *Proc. of the ICSLP'04*, Jeju, Korea.

1 Introduction

Prosody is an important part of the information structure of speech. An important prosodic gesture is stress. One possible definition given for stress describes it as the degree of loudness that a phoneme or syllable is pronounced with [1]. The importance of stress is attributed to the effect it has on the speech parameters. Earlier research has explored the effect of stress on duration, F0 and formant frequencies in Farsi (Persian) speech [2, 3]. One view on prosodic parameters is to try to make use of them in improving stressed automatic speech recognition (ASR) performance. It has already been shown that the use of second formant can lead to an average increase of 2.4% in stressed speech recognition rate, using baseline models [4]. In this work, we aim to further improve the stressed speech recognition performance using the prosodic parameters.

2 Evaluating the Effect of Stress on Cepstral Parameters

Research on Farsi has unveiled that stress affects such parameters as formant and fundamental frequencies [2, 3]. Research carried out on other languages agree to a great extent with these results. We tried to evaluate this effect on cepstral parameters. The effect of stress was evaluated on the first six cepstral parameters, namely C1 to C6. This effect was evaluated on both stressed and unstressed parts of the stressed sentences. We considered the changes in the cepstral parameters of all vowels due to stress and found that, in comparison to neutral (non-stressed) sentences, the scale of change was not consistent among the parameters. Therefore, finding a reliable relationship for modeling the changes in the cepstral parameters, according to stress, is rather difficult. Large cepstral parameter fluctuations in the stressed case could possibly justify the radical stressed speech recognition performance degradations when the baseline/unstressed models are used.

3 Frequency Warping for Stressed Speech Recognition

It is interesting to mention that the performance of our baseline continuous speech recognizer deteriorated from 66.67% for neutral speech to 43.52% for stressed speech. This emphasizes the importance of the effect of stress on speech parameters used in ASR. More details of our recognizer and the data used are given in Section 3.3.

© IEICE 2008
DOI: 10.1587/elex.5.187
Received January 01, 2008
Accepted February 06, 2008
Published March 25, 2008

3.1 Frequency Warping

MFCC parameters are found using filters applied to the short time speech spectrum. In compliance with the human hearing system, the well-known mel frequency scale is used, applying (1) to transform the linear frequency scale to mel scale [5].

$$Mel(f) = 2595 * \log_{10}\left(1 + \frac{f}{700}\right)$$

Investigations revealed that the ranges of change for the first three formant frequencies in the stressed case were [300 790], [1070 2320] and [2330 2930] Hz respectively [4]. As the filters used for MFCC calculations are placed in certain frequencies, their outputs will change due to the change in formant frequencies caused by stress. Taking into account the way stress changes the formant frequencies, we aim to warp the frequency scale in a way that such important frequency values, would be sent back to their original place (values for neutral speech). This can be viewed as a type of normalization to remove the effect of stress from the speech and is similar to the approach taken for vocal tract length normalization (VTLN) [6]. Obviously, this warping should be applied before applying (1).

3.2 Calculating the Frequency Warping Coefficients

A relationship is now needed between the neutral and stressed speech formant frequencies. In order to find this relationship, every stressed sentence and its neutral counterpart were taken. It should be noted that, at this stage, the stressed speech corpus was randomly divided into two sections with the size ratio of 2:1. The first of these was used to derive the warping equations and the second for recognition purposes. A scatter plot was then formed using these sets of data and neutral data. Figure 1 displays the overall scatter plots of the first three formants, drawn for all the voiced phonemes. They are derived using the stressed sections of the stressed training data and the neutral training data. According to Fig. 1, the closest resemblance to a line, in order to enable us apply a linear regression approximation, is seen in the second formant's scattered points, while for the other two formants no such relationship is observed. We divided the frequency range of [0 4000] Hz into three ranges of [0 900] Hz for F1, [900 2100] Hz for F2 and [2100 4000] Hz for F3. L2 in Fig. 1 represents the best line approximation for the scatter plot of the second formant frequency and features a slope equal to 0.964. To maintain frequency continuity, L1 and L3 lines were also used for F1 and F3 regions. Note that according to our experiments, the best results were obtained when L2 passed through the origin, therefore, L1 is the same as L2, while L3 is different (does not pass the origin). As we discussed earlier, stress does affect the formant frequencies of both the stressed and unstressed parts of the stressed utterances. Although we also found that the second formant in unstressed parts apparently changes in a less consistent manner in relation to that of the neutral speech, frequency warping will still be needed for both sections of the stressed utterances. The frequency warping for unstressed

DOI: 10.1587/elex.5.187
Received January 01, 2008
Accepted February 06, 2008
Published March 25, 2008

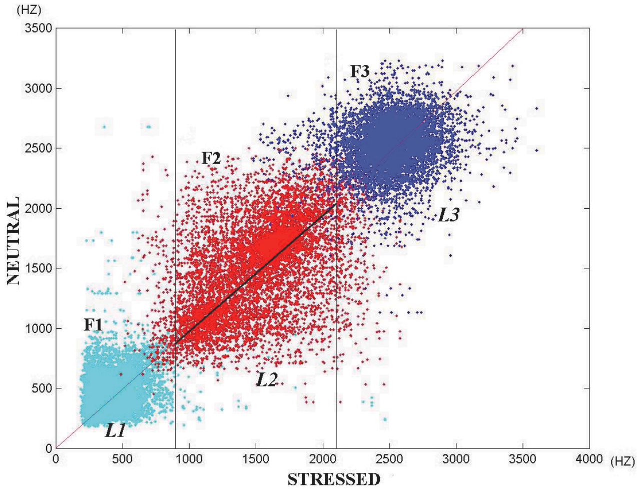

Fig. 1. Scatter plot of the first three formants in stressed and neutral train data.

parts were carried out similar to the stressed sections, but with a slope equal to 0.975 (for L1 and L2).

3.3 Stressed Speech Recognition Results Using Frequency Warping

The performance evaluation of our approach was carried out using FARSDAT continuous Farsi (Persian) speech corpus [7]. We chose 154 sentences from FARSDAT, which were believed to be suitable for applying stress. As the stress may be applied to more than one point in a certain sentence, an overall 468 stressed sentences were created with different points of stress. HTK [5] was used to create speaker-independent context-independent HMMs. The feature vectors consisted of 13 MFCC and log-energy coefficients plus their first and second order derivatives. The extraction of formants was carried out using standard software implementing the linear prediction technique [8].

A stressed/unstressed classifier, consisting of two GMMs, was trained and used to specify stressed and unstressed parts of the stressed sentences. The GMMs were trained using stressed and unstressed training data. This classifier demonstrated a performance of more than 95% in stressed speech classification [4]. The warping was applied to the frequency scale in the range specified for the second formant and the mel-scale filters redistributed. The newly formed MFCC parameters were then used to perform recognition on stressed database. Using the warped features and our baseline models, word recognition accuracy for stressed database increased to 44.68%, i.e. a relative improvement of about 2.7%. As shown in Fig. 1, linear regression was used to approximate second formant frequency variation due to stress and L1 and L3 were used to maintain continuity within the frequency scale in the ranges for F1 and F3. This probably inappropriate change of first and third formant frequencies may be counted as one reason for smaller than expected improvement in speech recognition accuracy.

© IEICE 2008
DOI: 10.1587/elex.5.187
Received January 01, 2008
Accepted February 06, 2008
Published March 25, 2008

3.4 Extended Warped-Frequency Feature Vector

It was shown that appending parameters such as the first three formants and their slope to the ordinary feature vector can lead to some improvements in the stressed speech recognition performance [9]. Using the baseline system with frequency-warped test features (called M0) and adding F1, F2 and F3 to the test feature vectors, three test configurations, namely M1, M2 and M3 were created. Three more test configurations, i.e. M4, M5 and M6, included slopes of the formant frequencies. Table I summarizes the word recognition results using the mentioned configurations. As mentioned, during the warping function calculations, the frequency scales of the first and third formants were also inevitably warped. Therefore, as shown in Table I, first and third formants are not useful during the recognition experiments and even lead to some performance deteriorations. This is also in accordance with our previous results that indicated that the second formant performed the best when added to the feature vector for experiments carried out on the aforementioned speech data [4]. The best performance was shown by the M2 configuration, where an absolute improvement of 4% was observed, in comparison to M0 configuration.

Table I. Recognition results for different feature vector configurations.

Models	M0	M1	M2	M3	M4	M5	M6
Accuracy	44.68	42.48	48.72	42.87	40.86	43.91	40.68

4 Conclusion

In this paper, a cepstral parameter compensation method for stressed speech recognition was proposed. In this approach, part of the frequency scale, in the range usually covered by the second formant, was warped using a transformation found from the more orderly changes in the second formant frequency values. This transformation was found based on linear regression. The warping was then applied to the frequency scale, prior to mel-scale filter calculations and the new cepstral parameters were found accordingly. The stressed speech recognition results obtained using the new feature parameters were superior to those of the ordinary parameters, giving an absolute overall performance improvement of about 1.3%. Appending second formant frequency to the warped-frequency feature vector led to further performance improvements of about 4% (absolute) in our experiments.

DOI: 10.1587/elex.5.187
Received January 01, 2008
Accepted February 06, 2008
Published March 25, 2008