5th International Conference on Spoken Language Processing (ICSLP 98)
Sydney, Australia
November 30 - December 4, 1998

Speech Feature Modeling for Robust Stressed Speech Recognition*

Sahar E. Bou-Ghazale and *John H. L. Hansen*
Robust Speech Processing Laboratory
Duke University, Box 90291, Durham, NC 27708-0291
sahar.bou-ghazale@rss.rockwell.com jhlh@ee.duke.edu

ABSTRACT

It is well known that the performance of speech recognition algorithms degrade in the presence of adverse environments where a speaker is under stress, emotion, or Lombard effect. This study evaluates the effectiveness of traditional features in recognition of speech under stress and formulates new features which are shown to improve stressed speech recognition. The focus is on formulating robust features which are less dependent on the speaking conditions rather than applying compensation or adaptation techniques. The stressed speaking styles considered are simulated angry and loud, Lombard effect speech, and noisy actual stressed speech from the SUSAS database (available on CD-ROM through NATO RSG.10 research group, and soon LDC). In addition, this study investigates the immunity of LP and FFT power spectrum to the presence of stress. Our results show that unlike FFT's immunity to noise, the LP power spectrum is more effective than the FFT to stress as well as to a combination of a noisy and stressful environment. Two alternative frequency partitioning methods (M-MFCC, ExpoLog) are proposed and compared with traditional MFCC features for stressed speech recognition. It is shown that the alternate filterbank frequency partitions are more effective for recognition of speech under both simulated and actual stressed conditions.

1. Introduction

It is well known that the performance of speech recognition systems degrade under the presence of stress. Stress in this context refers to speech produced under environmental, emotional, or workload stress. The stress conditions considered in this study include simulated angry and loud, Lombard effect conditions, and actual stressed speech all obtained from the SUSAS (Speech Under Simulated and Actual Stress) [4] database. The stress condition referred to as Lombard effect results when a speaker attempts to modify his or her speech production system while speaking in a noisy environment. To improve the performance of speech recognition algorithms under stress, previous studies have either adapted the recognizer to the input stressed speech during training [5] or compensated for the effect of stress during testing [1, 2, 6, 7]. These adaptations, however, require a preprocessing stage in order to model the statistics of the input stressed speech and to incorporate this knowledge in the recognition system which makes them specific to the stressed condition being addressed. A more general solution would be to improve the signal modeling. The ultimate goal would be to achieve a signal modeling framework or robust features which are immune to the speech variations due to stress. In addition, since the majority of features employed in current speech recognition systems are based on the FFT power spectrum due to its reported immunity to noise [8], this study will also investigate the immunity of the FFT power spectrum to stress, and contrasts its performance to the LP power spectrum. Finally, we emphasize that neutral speech data is used for all training evaluations (for each feature set). Round-robin open test evaluations are conducted for recognition of speech under various stressed conditions.

The remainder of this paper is organized as follows. The database employed in all of our evaluations is briefly summarized in Sec. 2. In Sec. 3, we investigate the recognition performance across individual frequency bands to determine frequency regions less affected by stress. Based on these results, we propose in Sec. 4, two new frequency scales which are less sensitive to the effects of stress as compared to the traditional mel-frequency scale. In Sec. 5, we compare the performance of LP and FFT power spectrum based features in the presence of stress. A final summary and a series of conclusions are drawn in Sec. 6.

2. Recognizer and Database

The speech data employed in this study is a subset of the SUSAS database [4]. All recognition evaluations are speaker-independent, and consider only male speakers. A 30-word HMM-based recognizer is formulated using a variable-state, left-to-right model, with 2 continuous mixtures per state. The HMM models are trained with neutral speech of eight speakers while a ninth speaker is left for open testing. The training and testing are done in a round robin scheme. In evaluating each of the neutral trained HMM models, a total of 2160 tokens are tested from the four speaking styles.

The last evaluation employs actual stressed speech from the SUSAS database which consists of speech produced

*This work was supported in part by a grant from the U.S. Air Force Research Laboratory/IFEC, Rome NY.

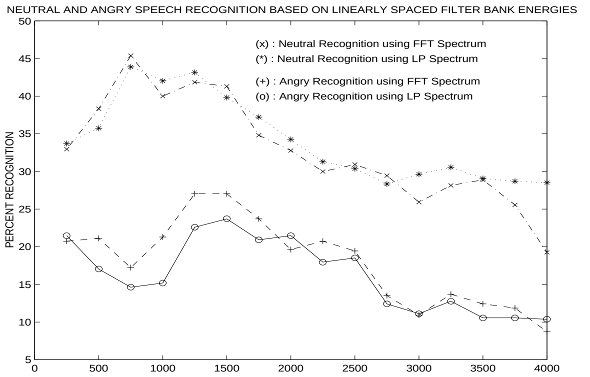

Figure 1: Recognition based on individual linearly-spaced filter bank output energies. Neutral trained models are tested with neutral and angry speech. The plot shows the results of both using an LP and an FFT power spectrum.

during the completion of two types of subject motion-fear tasks. The speakers produced speech while participating in two amusement park rides (e.g., a traditional roller-coaster ride and a free-fall ride consisting of a 130 ft vertical drop machine). These two rides were chosen in an attempt to simulate the sudden change in altitude or direction which could be experienced in an aircraft cockpit under emergency conditions [4].

3. Filter Bank Based Recognition

Here, we study the impact of stress on individual frequency bands. This is achieved by evaluating the recognition performance based on the log-energy output of a 16 uniformly spaced filter bank. The ultimate goal is to formulate a new frequency scale which is less sensitive to variations caused by stress without degrading the performance of neutral speech recognition. We note that a similar study proved to be successful in the formulation of a set of accent sensitive frequency features for accent classification [9]. A speaker-independent HMM model with variable state-duration is trained with neutral speech for each of the 16 frequency bands of a word. The neutral trained word models are tested with tokens of neutral and angry speech from SUSAS. The results shown in Fig. 1 are across 30 words spoken by nine speakers. The training and testing evaluations are done twice, once employing an FFT power spectrum, and a second employing a linear prediction spectrum.

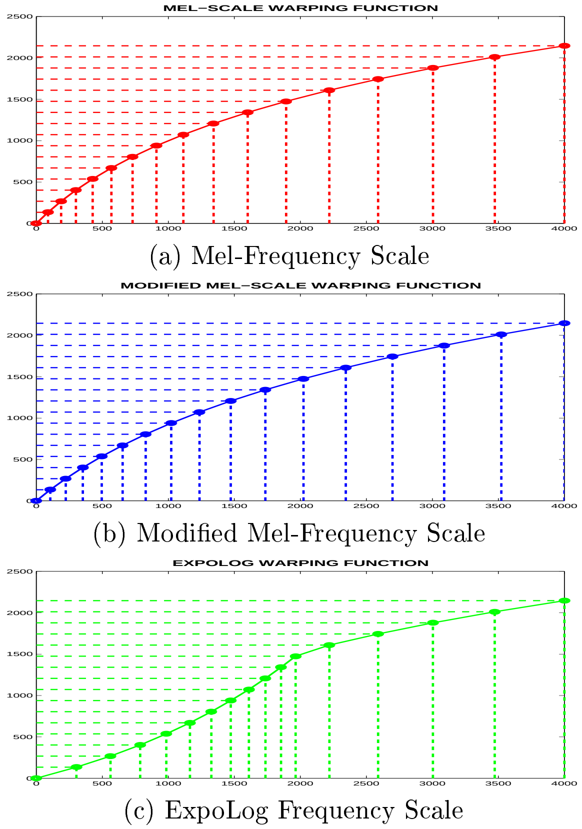

Figure 2: This graph shows three mapping functions, (a) mel-frequency, (b) modified mel-frequency, and (c) ExpoLog, employed to warp a linear scale in the frequency domain.

Fig. 1 shows that for both power spectral methods (LP and FFT), the highest recognition performance for neutral speech (top two lines) occurs around the first formant location (200 - 1000 Hz) while the highest recognition of angry speech (lower two lines) occurs in the neighborhood of the second formant location in the range of 1250 to 1750 Hz. This may be attributed to the observation that the second formant location is more closely related to tongue movement, and that the variations in tongue movement from neutral to stress conditions are less dramatic than other changes such as excitation for example. Therefore, since the second formant location experiences less variability under stress, it would be more reliable for stressed speech recognition. Recall that since a mel-scale is almost linear for frequencies below 1000 Hz and increases logarithmically above 1000 Hz, the contribution of the second formant is de-emphasized compared to the first formant. This attribute makes the mel-scale ideal for neutral speech recognition but not equally effective for angry speech recognition. In the following section, we introduce two new frequency scaling methods targeted at emphasizing the second formant in both neutral and stressed speech.

4. Newly Proposed Frequency Scales

To achieve the desired frequency analysis, we propose two new frequency partitions: one referred to as the modified mel-scale (M-MFCC), and the second is a combination of an exponential and a logarithmic function and is referred to as the ExpoLog scale. Both frequency scales along with the traditional mel-scale are given below:

$$\text{mel-scale} = 2595 \times \log(1 + \frac{f}{700}) \qquad (1)$$

$$\text{Modified mel-scale} = 3070 \times \log(1 + \frac{f}{1000}) \qquad (2)$$

$$\text{ExpoLog} = \begin{cases} 700 \times (10^{\frac{f}{3988}} - 1) & 0 \leq f \leq 2kHz \\ 2595 \times \log(1 + \frac{f}{700}) & 2 < f \leq 4kHz \end{cases} \qquad (3)$$

These three frequency warping functions are plotted in Fig. 2 for comparison. The y-axis represents the linear scale which is warped to the desired scale according to the mapping function. For the ExpoLog mapping, the filter banks are highly concentrated at mid frequencies while they are sparsely distributed at frequencies below 750 Hz and above 2000 Hz. The performance of the three warping functions is illustrated here using an FFT power spectrum. A comparison of their performance using an LP power spectrum will be discussed in the following section.

Fig. 3 shows results from an evaluation of the three frequency warping scales in obtaining cepstral parameters (MFCC, M-MFCC, ExpoLog). Recognition rates are shown for neutral models trained with static features and tested with speech from neutral and three stressed speaking conditions. When static features are employed for recognition, M-MFCC outperforms traditional MFCC by 4.45% for angry, 1.85% for loud, and 5.37% for Lombard effect. The performance of ExpoLog static features also outperforms the mel-scale, for all stress styles, with an average performance improvement of 4.77%. Note that for angry and loud speech recognition, ExpoLog exceeds MFCC by as much as 7.59% and 7.77%. These results clearly show that with a slight modification in the manner in which cepstral parameters are obtained, we can improve recognition performance in stressed speech conditions.

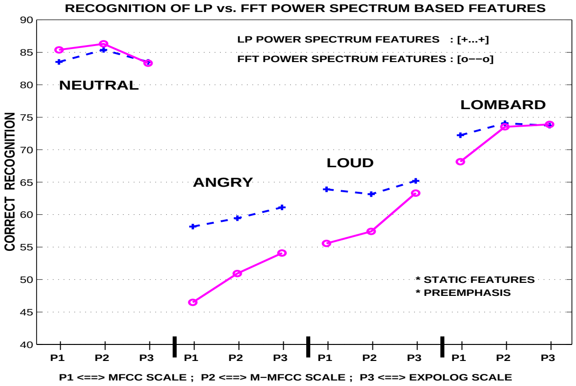

Figure 4: This graph compares the performance of FFT vs. LP power spectrum derived features of neutral trained models using static features. The performance of MFCC is compared to two different frequency scales.

5. FFT vs. LP Power Spectrum

In a recent survey by Picone [8] of contemporary recognition systems, it was established that FFT-based spectral parameters are preferred to LP-based parameters since they are believed to be more immune to the presence of noise. For this reason, a number of systems rely on the Fourier transform-based filter bank analysis. Our results based on individual frequency band output energies did not conclusively determine which power spectral estimation method (LP vs. FFT) would be more robust to stress. Here, we investigate the performance of both spectral estimation methods in stress using whole-word based models. We conducted two recognition evaluations using parameters derived from FFT and LP power spectral estimation methods. In addition, we performed an additional recognition evaluation employing actual stressed speech produced in a noisy environment in order to determine which power spectral estimation method would be more robust to the presence of both noise and stress. The noise in this case represents time varying mechanical and wind noise obtained from speech recorded during amusement park roller coaster rides.

Neutral Trained Models Employing Static & Dynamic MFCC, M-MFCC, and ExpoLog Features tested in noisy actual stressed conditions

Feature Set	LP Spectrum	FFT Spectrum
MFCC	36.72%	28.81%
M-MFCC	36.16%	25.42%
ExpoLog	37.29%	22.60%

Table 1: Performance of FFT and LP power spectrum based features in actual stressed noisy speech.

5.1. Noise-free simulated stress

Our results show that contrary to their noise immunity, FFT-based spectral parameters are not equally robust to the presence of stress. Fig. 4 compares the performance of LP and FFT power spectrum based features. The dotted line represents LP based recognition and the solid line represents FFT based recognition rates. The LP power spectrum performs significantly better than the FFT power spectrum when neutral trained models are tested with angry, loud, and Lombard effect speech. We also point out that modified MFCC (M-MFCC) and ExpoLog based features consistently outperformed MFCC parameters using both FFT and LP based spectra. In addition, LP derived ExpoLog produced the highest recognition rates across stressed styles using static features. Next, we consider extending the static features to include time derivatives and feature processing. Time derivatives or delta parameters were shown in a number of previous studies to greatly enhance the performance of stressed speech recognition. Having established the ExpoLog frequency scale as being superior to mel and modified-mel scales, we now consider time derivatives and parameter processing (liftering and cepstral mean normalization). Fig. 5 also compares both spectral methods. It shows the performance of ExpoLog static and dynamic features with parameter processing. Once again, the LP based features outperform FFT by an overall 3.94%. For angry speech recognition, the improvement in recognition is as much as +9.63%.

5.2. Actual noisy stressful conditions

This evaluation is intended to determine which power spectral estimation method is most effective when speech is subjected to a combination of noise and stress. The results, as summarized in Table 1, indicate that the LP-based features outperform the FFT-based features not only for noise-free simulated stress conditions but also for noisy actual stressed speech. We believe that the spectral smoothing inherent in the LP model provides a more overall

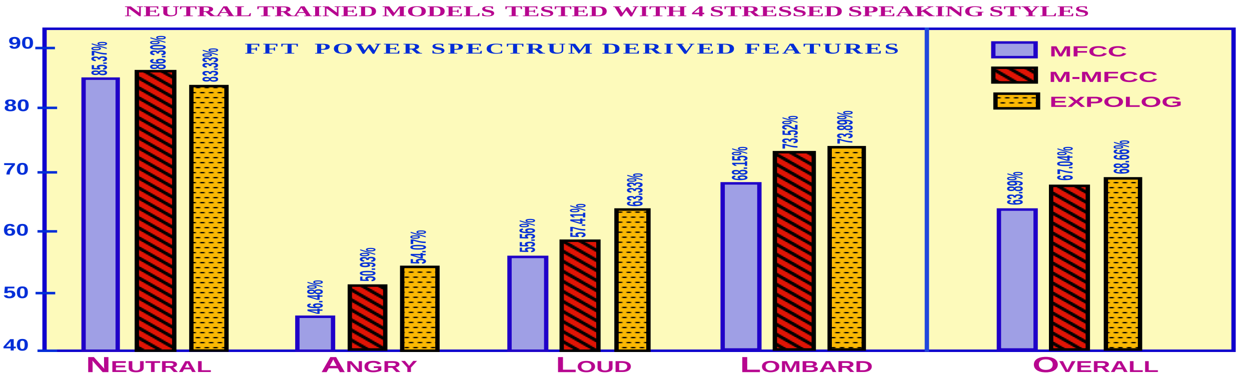

Figure 3: Performance of the static features of three different frequency mapping functions, MFCC, M-MFCC, and ExpoLog, based on an FFT power spectrum.

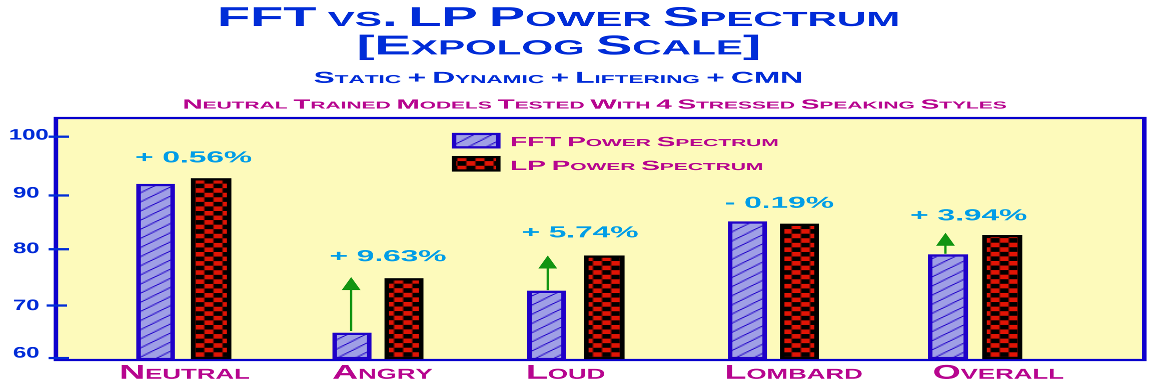

Figure 5: Performance of FFT and LP power spectrum based ExpoLog static and dynamic features.

smooth set of parameters capable of suppressing the fine variations caused by excitation changes (i.e., pitch structure) that occur under stressful conditions.

6. Conclusion

In this study, we formulated two new feature extraction methods, a modified mel-frequency scale (M-MFCC), and an exponential-logarithmic scale (ExpoLog) for improved stressed speech recognition. Both methods were shown to outperform the traditional mel-frequency cepstral parameters (MFCC) for recognition of speech under a variety of stressed styles.

This study also compared the performance of FFT-based features to LP-based power spectrum features in the presence of stress. Contrary to the FFT's immunity in noise, the FFT power spectrum was less robust to stress. Features based on the LP power spectrum outperformed the FFT-based features not only for noise-free simulated stress conditions but also for speech under actual noisy stressful conditions.

The final recommendation from this study is that for effective speech recognition performance in both neutral and stressed conditions, speech recognizers should (i) employ features derived from an LP as opposed to an FFT based power spectrum, and (ii) use a modified frequency partition such as M-MFCC or ExpoLog if possible.

7. REFERENCES

[1] Y. Chen, "Cepstral domain stress compensation for robust speech recognition," *ICASSP*, pp. 717–720, 1987.

[2] J.H.L. Hansen, *Analysis and Compensation of Stressed and Noisy Speech with Application to Robust Automatic Recognition.* Ph.D. thesis, Georgia Inst. of Tech., Atlanta, GA, July 1988.

[3] R.P. Lippmann, E.A. Martin, and D.B. Paul, "Multi-style training for robust isolated-word speech recognition," *ICASSP*, pp. 705–708, 1987.

[4] J.H.L. Hansen and S.E. Bou-Ghazale, "Getting started with SUSAS: A speech under simulated and actual stress database," *EUROSPEECH*, vol. 4, pp. 1743–1746, 1997.

[5] J.H.L. Hansen and S.E. Bou-Ghazale, "Robust speech recognition training via duration and spectral-based stress token generation," *IEEE-SAP* , vol. 3, pp. 415–421, 1995.

[6] B. Stanton, L. Jamieson, and G. Allen, "Robust recognition of loud and lombard speech in the fighter cockpit environment," *ICASSP*, pp. 675–678, 1989.

[7] J.H.L. Hansen and O.N. Bria, "Lombard effect compensation for robust automatic speech recognition in noise," *ICSLP*, pp. 1125–1128, 1990.

[8] J.W. Picone, "Signal modeling techniques in speech recognition," *Proc. IEEE*, vol. 81, pp. 1215–1247, Sept. 1993.

[9] L.M. Arslan and J.H.L. Hansen, "A study of temporal features and frequency characteristics in American English foreign accent," *JASA*, vol. 102, no. 1, pp. 28-40, July 1997.