Effects of Phoneme Characteristics on TEO Feature-based Automatic Stress Detection in Speech

Evan Ruzanski[1], John H.L. Hansen[1]
James Meyerhoff[2], George Saviolakis[2], Michael Koenig[2]

[1] Robust Speech Processing Group, Center for Spoken Language Research,
University of Colorado, Boulder, Colorado 80309-0594, USA
{ruzanski, jhlh}@cslr.colorado.edu

[2] Department of Neuroendocrinology, Division of Neuroscience,
Walter Reed Army Institute of Research (WRAIR), Silver Spring, Maryland

ABSTRACT

A major challenge of automatic speech recognition systems found in many areas of today's society is the ability to overcome natural phoneme conditions that potentially degrade performance. In this study, we discuss the effects of two critical phoneme characteristics, decreased vowel duration and mismatched vowel type, on the performance of automatic stress detection in speech using Teager Energy Operator features. We determine the scope and magnitude of these effects on stress detection performance and propose an algorithm to compensate for vowel type and duration shortening on stress detection performance using a composite phoneme decision scheme, which results in relative error reductions of **24%** and **39%** in the non-stress and stress conditions, respectively.

1. Introduction

Reliable stress detection can be used to enhance the performance, reliability, and robustness of speech recognition systems used in spoken dialog systems, cognitive task assessment, and spoken document retrieval, among other applications. Stress detection is also important in stand-alone applications, such as automatic assessment of stress levels of personnel in critical positions, such as pilots, air traffic controllers, and national defense personnel, allowing decisions and adjustments to be made regarding the suitability of such persons to adequately perform their duties and maintain the safety of others.

The problem of reliable stress recognition in speech recognition is compounded by the various characteristics of the speech samples used in the recognition scheme. It has been determined that vowels are an attractive class of phonemes to use as tokens in such recognition systems due to their definite quasi-periodic nature [1]. In this class of LDC (Linguistics Data Consortium[9]) phonemes, there are twenty different types of vowels to consider, each having different general characteristics.

In natural speaking situations, the duration, scope and breadth (e.g., due to position and coarticulation) of detectable types of the spoken vowels vary and these variations can degrade stress detection performance. Depending upon the type of speech (e.g. isolated word, spontaneous, read), phoneme durations can vary between words within the same sentence. Likewise, pronunciation of vowel types (e.g. front vs. middle vs. back vowel), can differ as well. In this paper, we address the effects of vowel duration and type mismatch using a test set of stressed and neutral (non-stressed) speech samples taken from a U.S. Army Soldier of the Month (SOM) board as used in [2]. This speech corpus consists of answers to questions posed in both stressful and non-stressful environments by English male speakers. The spoken answer to all questions is, "The answer to that question is 'no'". The stress and neutral conditions have been verified by biometric data readings as summarized in [2]. In [2], it was shown that the mean heart increased by 34.2%, systolic blood pressure increased by 33.0%, and diastolic blood pressure increased by 22.5% from neutral to stress conditions.

The neutral speech data was collected over six separate recording sessions over a time period from 1 hour to seven days before (i.e. sets "a", "b", and "c") to 1 hour to seven days after (i.e. sets "e", "f", and "g") and on the SOM board (i.e. set "d").

The performance degradation due to discrepancies in the vowel type and duration characteristics can be corrected for by considering the collection of vowels within the sentence. To this end, we decide the condition (i.e. stressed or not stressed), of the speaker by employing a weighted majority decision rule based upon the decisions of each vowel within the sentence. The decisions are weighted by the absolute difference between two Hidden Markov Model (HMM) scores. This Manhattan distance metric gives a "confidence measure" of the condition of the token (i.e. a greater distance indicates greater surety in the correct decision made by the HMM). As we will show, this algorithm yields a substantial performance enhancement for automatic stress detection under the actual military stress test conditions described above and presents a significant step towards reliable stress detection for spontaneous, unrestricted, conversational speech.

2. Teager Energy Operator-based stress classification

Historically, most approaches to speech modeling have taken a linear plane wave point-of-view. While features derived from such analysis can be effective for speech coding and recognition, they are clearly removed from physical speech modeling. Teager [3, 4] did extensive research on nonlinear speech modeling and pioneered the importance of analyzing speech signals from an energy point-of-view. He devised a simple nonlinear energy-tracking operator that can model the airflow through the vocal tract, shown for discrete-time signals as follows:

$$\Psi[x(n)] = x^2(n) - x(n+1)x(n-1), \qquad (1)$$

where $\Psi[\cdot]$ is the Teager Energy Operator (TEO). Kaiser first systematically introduced the TEO [5, 6].

Authorized licensed use limited to: Univ of Texas at Dallas. Downloaded on August 31, 2009 at 14:46 from IEEE Xplore. Restrictions apply.

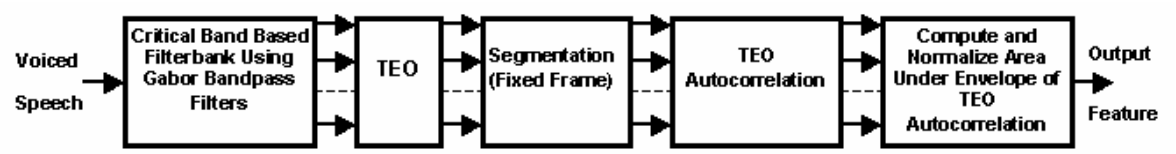

Fig. 1: Feature Extraction Flow Diagram

The Teager Energy Operator, Critical Band, Autocorrelation Envelope (TEO-CB-Auto-Env) parameterization feature has been shown to reflect variations in excitation \under stressful conditions [1]. A speech signal's fundamental frequency will change and hence the distribution pattern of pitch harmonics across critical bands will be different than for speech under non-stressful conditions [1]. This finer frequency resolution comes from the partitioning of the entire audible frequency range into many critical bands [7, 8].

The TEO-CB-Auto-Env feature used in this study is extracted through a process shown in the flow diagram of Fig. 1 and illustrated mathematically using critical bandpass filters (BPF) as,

$$u_j(n) = s(n) * g_j(n),$$

$$\Psi_j(n) = \Psi[u_j(n)] = u_j^2(n) - u_j(n-1)u_j(n+1),$$

$$R_{\Psi_j^{(i)}}(k) = \sum_{n=1}^{N-k} \Psi_j^{(i)}(n)\Psi_j^{(i)}(n+k),$$

where,

$g_j(n), j = 1,2,3,...,16,$ is the BPF impulse response,

$u_j(n), j = 1,2,3,...,16,$ is the output of each BPF,

"*" is the convolution operator,

$R_{\Psi_j^{(i)}(n)}(k)$ is the autocorrelation function of the

i^{th} frame of the TEO profile from the j^{th} critical band,

$\Psi_j^{(i)}(n), j = 1,2,...,M,$ and N is the frame length.

While this feature has been shown to be effective for stress classification [1, 2], there are some fundamental questions that should be addressed. This study addresses the following two questions.

First, effective stress classification is possible using voiced speech data from vowels with the TEO-CB-AutoEnv where critical bands are weighted based on neutral/stress detection performance from a development test set [2]. What impact does reducing the phoneme duration have on stress detection? We know vowel duration is reduced in multi-syllable words for syllables with lower lexical stress levels.

Second, since the TEO-CB-AutoEnv employs an autocorrelation envelope analysis on the input speech, it should be less sensitive to phoneme spectral differences in stress detection than spectral features such as MFCCs (mel-Frequency cepstral coefficients) ([1] proved this point). What impact is there on stress detection performance if TEO-CB-AutoEnv neutral/stress models are trained with data from one set of phonemes and tested with TEO features obtained from other phonemes?

3. Phoneme duration effects on stress detection performance

The 8-kHz digitized speech data from the Soldier of the Month (SOM) Board was processed for isolation of the vowel /OW/ from each of the forty-two sentences from each of the original corpus of six speakers. A total of two hundred fifty-two extractions were performed. A Hidden Markov Model (HMM)-based speech alignment program was used for such extractions by cropping the selected vowel from the remainder of the test sentence.

Each extracted sample of the vowel /OW/ was verified to be valid, and manual front- or back-end processing was done to ensure a complete and accurate representation of the vowel /OW/ (e.g., remove audible instances of the phoneme /N/ from the complete word "NO", and removal of trailing silence critical for phoneme duration testing). These extracted vowels were considered to be the "100%-duration" vowels.

The probability density functions (PDF) depicting the time duration probability distributions for the neutral and stress sets are shown in Fig. 2, with mean values of 214 ms and 203 ms and standard deviation values of 63 ms and 63 ms for the neutral and stress set, respectively. The similar mean and standard deviation values suggest the duration of the vowel /OW/ is not affected by stress content in the SOM speech data.

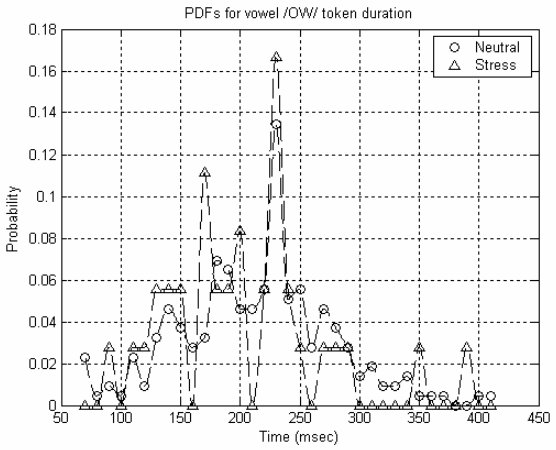

Fig. 2: Probability mass function of vowel /OW/ tokens

Each of the "100%-duration" vowel samples was truncated (by taking samples from the back-end) to specified durations of 80, 60, 40, and 20% of the 100% vowel duration length. The Teager Energy Operator (TEO) feature was extracted on a frame-by-frame basis from each vowel instance in each of these five duration sets.

Based on past experimental data, the frame lengths for the feature extraction were fixed to two hundred samples [2]. The amount of shift of this 200-sample frame was set to the values of one hundred samples and twenty-five samples, constant across the full set of vowel tokens. The hypothesis was that in setting the frame shift to a lower value, more TEO feature values could be computed in a given duration sample. This fact will become critical in the HMM testing and scoring of the lower (i.e. 40 and 20%), vowel durations.

Previous research has shown that a selective band-weighting scheme significantly improves stress detection performance [2].

Authorized licensed use limited to: Univ of Texas at Dallas. Downloaded on August 31, 2009 at 14:46 from IEEE Xplore. Restrictions apply.

For the present study, the critical bands in the TEO extracted features were equally weighted (i.e. baseline tests were performed). The band-weighting scheme was not used in the duration testing. The goal of this experiment is to determine the effect of vowel duration on stress detection performance and so we have deferred introducing the variable band-weighting scheme, as that would introduce and additional analysis dimension. This explains the relatively high error percentages for the experiments vs. those reported in [2].

The extracted features from the "100%-duration" vowel /OW/ were used to train the HMMs used for the "round robin" testing. The procedure is summarized as follows: five of the six vowel token sets from all speakers were used to train two separate HMM models, one using the neutral tokens (i.e. token sets "a", "b", "c", "e", "f", "g") and one using the stress tokens (i.e. token set "d"). This gives a total of twelve 3-state, 2- Gaussian mixture HMMs, six each for neutral and stress.

The six HMMs come from the fact that five of the six tokens are used in a "round robin" to train the HMMs (i.e. token sets 1, 2, 3, 4, 5 produced one model representing the neutral condition using the neutral tokens described above and one for the stress condition using the stress token set). Another HMM set, neutral and stress, was similarly trained using token sets 1, 2, 3, 4, 6, another using token sets 1, 2, 3, 5, 6, and so forth. The reader should note that only the "100%-duration" vowels were used in the HMM training and testing was performed with 100% and progressively shorter duration vowel test material. Once the HMMs are trained accordingly, the single token that was left out of the training set was submitted to both HMMs. The HMMs were scored using existing code. For example, if a neutral token was submitted to the neutral- and stress-trained HMMs, and the stress-trained HMM produced a higher log likelihood score, a detection error was recorded.

Fig. 3 illustrates the effect of vowel duration on the performance of the TEO stress detection scheme using TEO features extracted at frame shift rate of twenty-five samples.

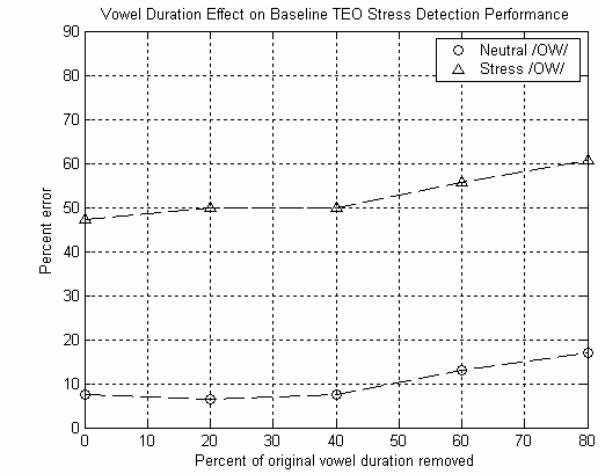

Fig. 3: Vowel duration effect on baseline TEO stress detection performance

From the examination of Fig. 3, it is apparent that stress detection performance remains nearly constant up to removal of between 40 and 60% of the original vowel duration for both neutral and stress speech. This suggests the mean duration values for speech tokens to yield desirable stress detection performance are approximately 85.6-128.4ms for neutral and 81.2-121.8ms for stress, based on mean durations shown in Fig. 1 and therefore a phoneme duration threshold should be set for effective stress detection.

A relatively large number of TEO feature values is necessary to ensure that proper amounts of feature data are used to train HMMs. In the 100-sample frame shift case, at 40% original vowel duration, 32/216 neutral test tokens did not contain enough information to be scored properly by the HMMs. At 20% vowel duration, this number increased to 105/216. In the 25-sample frame shift case, at 40% vowel duration, all tokens were properly scored by the HMM and at 20% vowel duration, only 16/216 of the tokens did not contain enough information to be properly scored. No further analysis of the frame shift was deemed necessary, as at 20% vowel duration in the 25-sample frame shift case, only 7.4% of the tokens were not useful for the test. The 25-sample frame shift will be used for the remainder of the TEO feature extractions in this study.

4. Analysis of vowel type difference on stress detection performance

As with the duration experiment above, we isolate the vowel type in the following way. First, we employ the baseline stress detection scheme to eliminate the effects of vowel type on the critical band weighting. Next, we fix the duration of each token in the set to 50ms. We note that since the word "no" is mono-syllabic and a keyword in the sentence response, the /ow/ was typically longer in duration than other the other vowels. We choose this value to allow for a reasonable number of tokens across the vowel types thus allowing a broad test set among vowel types. Selecting this duration also allows for a small amount of TEO feature data to be submitted for testing and illustrates the effectiveness of our majority rule stress detection scheme introduced later.

The vowel types we use for the type testing include, /AE/, /AX/ /IY/ and /OW/. The collection of full-length vowels among these vowel types across the sentences in the test set was used to train a 3-mixture, 3-state multi-style HMM. We chose the multi-style HMM to account for the various vowel types in one model and introduced an additional Gaussian mixture for better resolution between vowel types. We then submit each vowel type set for testing. The results are summarized in Table 1.

	Neutral/Stress Classification Error (%)		
Test Vowel Type		**Neutral**	**Stress**
	/AE/	**31.25**	**37.14**
	/AX/	**44.17**	**54.28**
	/IY/	**33.90**	**38.89**
	/OW/	**33.1**	**50.00**
Overall Performance	***μ***	***35.60***	***45.06***
	σ	***5.82***	***8.37***

Table 1: Multi-style HMM stress detection performance across vowels types

The results in Table 1, namely the standard deviation values, σ, show small variability in stress detection performance across vowel types when using the multi-style HMM trained on the full-duration vowel tokens, with some vowel types performing slightly better under neutral and stress conditions. We note that test materials were 50ms in duration, which would include at least one test block for most vowel sections extracted (i.e., from Fig. 1, on

Authorized licensed use limited to: Univ of Texas at Dallas. Downloaded on August 31, 2009 at 14:46 from IEEE Xplore. Restrictions apply.

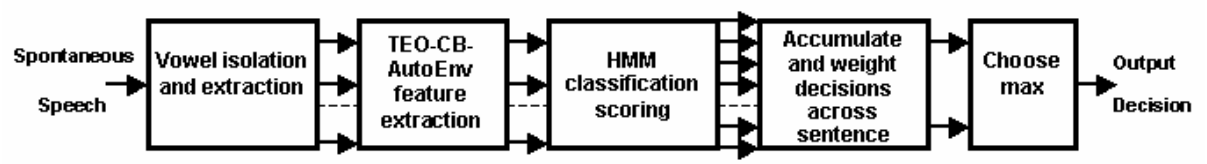

Fig. 4: Flow Diagram for Composite Neutral/Stress Detection Scheme

average there would be between 2-8 test blocks available from a single /ow/ vowel section). We see from the results in Table 1, the error rates average μ = 35.60 percent for neutral and μ = 45.06 percent for stress. By using a composite detection scheme, these error rates are reduced significantly.

5. Novel scheme for improving stress detection across vowel types with decreased durations

The idea for improving stress detection performance in this test environment lies in the exploitation of multiple vowel types per speaker utterance and making a classification decision based upon a majority rule over individual decisions made on each vowel in the sentence separately. This procedure is illustrated in Fig. 4.

The test results using the HMM training scheme employed in the previous section, with an overall decision based on the composite weighted majority decision rule scheme results in significant performance gains. Table 2 shows improvement over the average error rates produced by using the individual phoneme decisions.

	Percent error	
	Neutral	**Stress**
Average error rate for individual vowel decision scheme	**35.60%**	**45.06%**
Error rate for composite phoneme decision scheme	**27.06%**	**27.78%**

Table 2: Performance Comparison of Individual Vowel vs. Composite Vowel Decision Schemes

For the test set used in these experiments, there was an average of 4.22 and 4.50 vowel observations per sentence for the neutral and stress condition sentences, respectively. Therefore, on the average the improvement in Table 2 comes from employing a decision based on four phonemes versus one.

6. Summary and Conclusion

In this paper, we have shown the effects of phoneme duration and type on automatic Teager Energy Operator feature-based stress detection performance. It was shown that both phoneme duration and type mismatch affect the stress detection performance. In the case of vowel duration, shortening the vowel duration was shown to adversely affect stress detection performance if the vowel duration is less than a threshold of about 50% of the original duration in the case of the vowel /OW/, or at about 85-128 ms. In the case of vowel type, it was shown that stress detection performance varies among and between the different vowel types whose durations were sufficiently long to use the HMM-based stress detection scheme.

It was also shown that in this test protocol, where a speaker's utterance contains multiple vowel types, using a composite decision process by applying a weighted majority decision rule to the individual decisions with equal (i.e. baseline) band-weighting provides relative error rate reductions of 24% and 39% in the neutral and stress conditions, respectively. The final composite stress detection scheme achieves overall average error rates of approximately 27% for neutral/stress detection, where the vowel durations are 50 ms. This suggests that the composite vowel decision algorithm outlined here yields a substantial performance enhancement for automatic stress detection under the test conditions described above and presents a significant step towards reliable stress detection for spontaneous, unrestricted, conversational speech.

This scheme may also be applied to stress detection in individual phonemes of adequate duration. The phoneme could be segmented into shorter duration segments and the composite decision scheme applied to each segment to arrive at an overall decision for the complete phoneme.

The error rates presented may be further improved by employing a soft decision rule algorithm, based upon a criteria associated with the characteristics of specific vowels, such as weighting decisions made on longer duration vowels more heavily than those from shorter duration vowels.

Another important characteristic regarding the robustness of automatic stress recognition systems is the issue of speaker dependence. Further research in this area would complement the findings presented here.

7. Acknowledgements

This work was supported by US Army Medical Research and Materiel Command (MRMC) under direction from WRAIR, grant W23RYX-4118-N601. Any opinions, findings, and conclusions expressed are those of the authors and do not necessarily reflect the views of U.S. Army.

REFERENCES

[1] G. Zhou, J.H.L. Hansen, and J.F. Kaiser, "Nonlinear feature-based classification of speech under stress", *IEEE Trans. Speech & Audio Process.,* **9** (3):201-216, Mar. 2001.

[2] M. Rahurkar, J.H.L. Hansen, et.al., "Frequency Distribution Based Weighted Sub-band Approach for Classification of Emotional/Stressful Content in Speech", EUROSPEECH-2003/INTERSPEECH-2003, Switzerland, 2003.

[3] H. Teager, "Some Observations on Oral Air Flow During Phonation", *IEEE Trans. Acoustics, Speech & Signal Proc.,* **28**(5):599-601, 1990.

[4] H. Teager, S. Teager, "Evidence for Nonlinear Production Mechanisms in the Vocal Tract", *Speech Production and Speech Modeling,* NATO Advanced Study Institute, vol. 55, Kluwer Academic Pub., pp. 214-261, 1990.

[5] J. F. Kaiser, "On a Simple Algorithm to Calculate 'Energy' of a Signal", ICASSP-90, pp. 381-384, 1990.

[6] J. F. Kaiser, "On Teager's Energy Algorithm, its Generalization to Continuous Signal", in *Proc. 4th IEEE Digital Signal Processing Workshop,* Sept. 1990.

[7] B. Scharf, "Critical Bands", *Foundations of Modern Auditory Theory,* J.V. Tobias(Ed), Academic Press, Vol. 1, pp. 157-202, 1970.

[8] W. A. Yost, "Fundamentals of Hearing", 3rd Edition, Academic Press, pp. 153-167, 1994.

[9] LDC: http://www.ldc.upenn.edu/

Authorized licensed use limited to: Univ of Texas at Dallas. Downloaded on August 31, 2009 at 14:46 from IEEE Xplore. Restrictions apply.