

FREQUENCY DISTRIBUTION BASED WEIGHTED SUB-BAND APPROACH FOR CLASSIFICATION OF EMOTIONAL/STRESSFUL CONTENT IN SPEECH*

Mandar A. Rahurkar and John H.L. Hansen

Robust Speech Processing Group,
Center for Spoken Language Research, University of Colorado
Boulder, CO-80303
{rahurkar, jhlh}@cslr.colorado.edu
Web: <http://cslr.colorado.edu>

ABSTRACT

In this paper we explore the use of nonlinear Teager Energy Operator based features derived from multi-resolution sub-band analysis for classification of emotional/stressful speech. We propose a novel scheme for automatic sub-band weighting in an effort towards developing a generic algorithm for understanding emotion or stress in speech. We evaluate the proposed algorithm using a corpus of audio material from a military stressful Soldier of the Quarter Board evaluation panel. We establish classification performance of emotional/stressful speech using an open speaker set with open test tokens. With the new frequency distribution based scheme, we obtain a relative detection error reduction of **81.3%** in stress speech, and a **75.4%** relative detection rate reduction in neutral speech detection error rate. The results suggest a important step forward in establishing an effective processing scheme for developing generic models of neutral and emotional speech.

1. INTRODUCTION

The problem of detecting emotion in speech has been the subject of a number of studies [1, 2, 3]. Much of the current effort on detecting emotions has been aimed at detecting emotion for improving the robustness of speech recognition algorithms. However, depending on the type of emotion or task induced stress condition, reliable detection, even in noise free environments, continues to be a challenging task. Reliable stress detection requires that a speaker change their neutral speech production process in a consistent manner so that extracted features can detect and perhaps quantify the change. However, there is significant variability in how different speakers convey stress or emotion. A previous study on stress speech classification [1] resulted in the formation of a nonlinear Teager Energy Operator (TEO) based feature

TEO-CB-AutoEnv. That study considered a critical band partition where autocorrelation area coefficients were determined for each filter band. Our recent research[10] has confirmed that the presence of stress in speech, for voiced speech phonemes, is not uniformly distributed across frequency. Initially, we selected four frequency sub-bands that's showed the lowest classification error, to construct a simple weighted scoring scheme. However, it stands to reason that there should be a more formal way to determine the sensitive sub-bands that are consistent for stress/neutral classification. This is especially important if we consider new speakers within the stress classification scenario. Therefore, our goal is to formulate an algorithm for automatic sub-band selection and weighting to obtain the best system for emotional speech classification. This algorithm would take us a step closer to a generic processing sequence for obtaining models of new emotions in a fixed speaker environment of new input speakers for voice telephony or interactive systems.

Another motivation for addressing emotional/stressful speech classification is to better understand physical speech modeling and associated speech feature estimation procedures. The idea of using TEO instead of the commonly used mean squared energy, is to take advantage of the modulation energy tracking capability of this feature. This property is especially useful here because we believe that when a person is under stress, physiological changes occur in their vocal fold movement, which modulates airflow in the vocal tract. These changes in the airflow properties are perceived by listener as emotional content.

Such an emotion classification system could not only be used to increase robustness in speech recognition it could also contribute to improved systems for to

*This research was funded by DARPA through SPAWAR under Grant No. N66001-00-2-8906

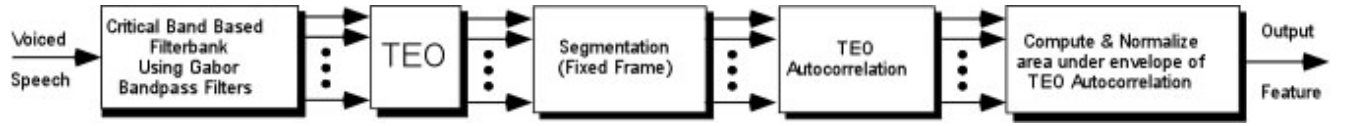


Figure 1: Feature extraction Flow Diagram

handhelds, interactive books, and intelligent dialogue systems. These are some of the many areas where such a system can play a vital role.

2. TEAGER ENERGY OPERATOR

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 did extensive research on nonlinear speech modeling and pioneered the importance of analyzing speech signals from an energy point of view[5,6]. He devised a simple nonlinear, energy tracking operator, for a continuous time signal $x(t)$ as follows:

$$\varphi_c[x(t)] = [\dot{x}(t)]^2 - x(t)\ddot{x}(t), \quad (1)$$

and for a discrete-time signal $x(n)$ as:

$$\varphi[x(n)] = x^2 - x(n+1)x(n-1), \quad (2)$$

where $\varphi[\cdot]$ is the Teager Energy Operator (TEO). These operators were first introduced systematically by Kaiser[7,8].

It has been observed[1] that under stressful conditions, a speech signal fundamental frequency will change and hence the distribution pattern of pitch harmonics across critical bands will be different then for speech under neutral conditions. Therefore, for finer resolution of frequencies, the entire audible frequency range can be partitioned into many critical bands. Each critical band possesses a narrow bandwidth, (i.e., typically 100-400Hz), thus making this new feature independent of the accuracy of median F0 estimation. This is essential as reliable pitch estimation in emotional speech is difficult, since pitch can increase by more than 200 percent in some high stress situations [11].

2.1. TEO-CB-AutoEnv : Critical Band Based TEO Autocorrelation Envelope

We can summarize the feature extraction procedure mathematically as follows using bandpass filters (BPF) centered at critical band frequency locations,

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

$$\varphi_j(n) = \varphi[u_j(n)] = u_j^2(n) - u_j(n-1)u_j(n+1),$$

$$R_{\varphi_j^{(i)}(n)}(k) = \sum \varphi_j^{(i)}(n)\varphi_j^{(i)}(n+k),$$

where,

$g_j(n)$, $j = 1, 2, 3, \dots, 17$, is the BPF filter response,

$u_j(n)$, $j = 1, 2, 3, \dots, 17$, is the output of each BPF,

$R_{\varphi_j^{(i)}(n)}(k)$ = Autocorrelation function of the i th frame of the TEO profile from the j th critical band, $\varphi_j^{(i)}(n)$, and, N = Frame length.

Fig.1 shows a flow diagram of the feature extraction process. The TEO-CB-AutoEnv feature[1] has been shown to reflect variations in excitation characteristics including pitch harmonics, due to its finer frequency resolution. However, we believe that the variation in excitation structure is not uniform across all the bands. In our previous work [10], we proposed a new weighted scheme which supported our hypothesis. Moreover, we also showed that specific frequency bands are more sensitive to stress and some to neutral. However, the bands were determined by testing over the same speakers in the training set, though the test speech tokens were different from those in the training set.

3. EXPERIMENTAL SETUP

3.1 Soldier of the Quarter Board (SOQ) Speech Corpus

A speech under stress corpus was collected at the Walter Reed Army Institute of Research. The speech corpus was constructed in WRAIR Soldier of the Quarter paradigm[11,12], by recording the spoken response of 6 individual soldiers to questions in a neutral setting, and while seated in front of a seven person military evaluation board (all board members had military rank above the soldier who faced the panel). The SOQ board is a training exercise and a competition used to prepare soldiers for actual promotion boards. Subjects in this study were candidates in the competition who volunteered to be studied after giving informed consent. Table 1 summarizes average speaker conditions for 6 speakers and 7 speech data collection phases before "Day of Board (DOB)" (A, B, C), during DOB (D), and after DOB (E,F,G). Changes in mean heart rate(HR), blood pressure(sBP, dBP) and pitch(F0) all confirm a change in speaker state between {A,B,C,E,F,G} and D. Further blood chemical analysis also confirmed that speakers under DOB (D) stress condition were in fact stressed.

Summary Of Mean Biometrics For SOQ Subjects					
Measure	A B -7 Day	C -20 min	D Board	E +20 min	A B +7 Day
HR	70.3	70.8	93.2	69.5	67.2
sBP	118	146	178	154	117
dBP	77.5	74.8	89.7	71.2	69.5
F0	103.4	102.7	136.9	104.3	103.1

Table 1: HR - heart rate (in beats per minute), SBP -Systolic blood pressure in mm, dBP – Dystolic blood pressure in mm, F0 – Fundamental frequency in Hz.

Results confirm a significant shift in biometric measures from the assumed neutral conditions (A,B,C),(E,F,G), versus the assumed stress condition (D). Each soldier was asked to answer all questions by responding "the answer to this question is NO". Each speaker was asked the same set of 6 different militarily-relevant questions on seven occasions. For our evaluations, we focused our analysis on the word 'NO'.

4. EVALUATIONS

For all evaluations, we used a Hidden Markov Model (HMM) classifier, with three states and two Gaussian mixtures. The evaluations were performed on the speech sound 'no' extracted from the corpus. Six speakers were used for evaluations

4.1 Baseline Evaluation

For baseline evaluation we divided the available corpus into three sets: training set, development test-set and open testing set. Four speakers were used for training, one for development, and one speaker was set aside for testing thus allowing us to carry out open speaker evaluation. Since the corpus is not large enough to allow the luxury of independent sets, we did a round-robin amongst six speakers wherein each speaker was tested against combination of remaining five speakers. Thus, each of the five speakers acted as a development set. Baseline results were averaged over all 30 of these evaluations.

4.2 HMM for Frequency Band Analysis

For frequency band analysis, a second HMM classification system was trained with a front-end feature made up of the TEO-CB-AutoEnv of each individual band. Thus, we have an independent HMM system for each band. A separate Neutral and Stress model was therefore constructed for every band. We therefore have thirty-four single band models, seventeen neutral and seventeen stress respectively. Evaluations were performed in the same way as in baseline evaluation by using three different sets for training, development and testing. Development sets were used to determine the band weights using the new scheme.

5. NEW FREQUENCY DISTRIBUTION BASED BAND WEIGHING SCHEME

We realize that selecting a small set of bands for classifying emotional speech may not result in a consistent classifier for all speakers. Hence, we propose here to use all the bands and not only those with low error percentage. However, if all bands are weighted equally, the bands which degrade the performance may offset the performance of sensitive bands. This is especially true in circumstances where the difference between stress and neutral speech is very subtle. In order to solve this problem, we developed a novel automatic weighting scheme for bands where weights are assigned based on training and development test sets.

		Neutral Speech							
SPKRS	166n	168n	169n	170n	171n	2168n	F.D	Weights	
Band 1	4	3	4	1	2	4	18	0.12	
Band 2	4	5	4	5	5	5	28	0.18667	
Band 3	1	2	0	0	1	0	4	0.02667	
Band 4	0	2	1	1	0	1	5	0.03333	
Band 5	1	2	0	1	0	1	5	0.03333	
Band 6	1	0	1	0	1	1	4	0.02667	
Band 7	5	4	5	5	5	3	27	0.18	
Band 8	1	0	0	0	0	1	2	0.01333	
Band 9	0	0	2	2	5	1	10	0.06667	
Band 10	1	1	1	0	1	1	5	0.03333	
Band 11	3	2	1	3	3	4	16	0.10667	
Band 12	0	0	0	0	0	0	0	0	
Band 13	0	0	0	0	0	0	0	0	
Band 14	3	2	4	5	2	1	17	0.11333	
Band 15	0	0	0	0	0	1	1	0.00667	
Band 16	0	0	0	0	0	0	0	0	
Band 17	1	2	2	2	0	1	8	0.05333	
						SUM	150	1	
		Stress Speech							
SPKRS	166s	168s	169s	170s	171s	2168s	F.D	Weights	
Band 1	0	0	0	2	0	0	2	0.01667	
Band 2	1	0	1	0	0	0	2	0.01667	
Band 3	2	1	0	2	3	0	8	0.06667	
Band 4	2	0	2	1	3	4	12	0.1	
Band 5	3	2	3	3	3	2	16	0.13333	
Band 6	1	3	2	1	2	1	10	0.08333	
Band 7	0	0	0	0	0	0	0	0	
Band 8	1	0	2	0	2	2	7	0.05833	
Band 9	0	2	0	1	0	0	3	0.025	
Band 10	0	0	2	0	0	0	2	0.01667	
Band 11	0	1	0	0	0	0	1	0.00833	
Band 12	2	2	1	3	1	1	10	0.08333	
Band 13	5	3	4	3	2	1	18	0.15	
Band 14	0	0	0	0	0	0	0	0	
Band 15	3	3	2	3	2	3	16	0.13333	
Band 16	0	3	1	0	2	4	10	0.08333	
Band 17	0	0	0	1	0	2	3	0.025	
						SUM	120	1	

Table 2: Frequency Distribution table for determination of band weights. FD = Frequency Distribution

The scheme, which is very elegant, is also computationally simple. For neutral speech, the frequency distribution for each band that occurs in the top 5 performing bands in each evaluation was computed. These distributions were summed up across all speakers so as to have generic weights. Next each frequency distribution was divided by the sum, thus not discriminating between any of the 17 bands. As can be seen in Table 2, all the weights sum to one. For stressed speech, the frequency distribution was computed by selecting the top four bands in speaker test in development set. The selection of number of bands to use was determined empirically. The last column shows the computed weights. After the weights have been computed, we test these weights using the scoring scheme which we proposed earlier[10]. Thus, we are now looking at a combination of weighted scores instead of just choosing the higher score. The weighted score is calculated below using Eqn.3,

$$Score = \sum_{n=1}^{17} W_{(n)}SSB(n) - \sum_{n=1}^{17} W_{(n)}NSB(n) \quad (3)$$

where,

SSB (n) = Stress Score of Sub-Band n,

NSB (n) = Neutral Score of Sub-Band n,

$W_{(n)}$ = Weight for band n, n=1, 2...17.

The results of evaluations using new weighing scheme are shown in Table 3. Using the entire TEO-CB-AutoEnv area feature vector, baseline stress and neutral error rates are 69.72% and 16.20% respectively. Using our new Automatic Subband weighting scheme the percentage stress and neutral rates dropped down to 13.06% and 3.98% respectively. This corresponds to a relative reduction of **81.3%** in stress speech detection error rate, and a **75.4%** percent reduction in neutral speech detection rate.

EVALUATION	%NEUTRAL ERROR	% STRESS ERROR
Baseline	16.20%	69.72%
Auto-Weight SB	3.98%	13.06%

Table 3: Evaluation using New Detection Scheme.

6. CONCLUSION

In this study, a new frequency distribution based Sub-band weighting method for stressed speech recognition is proposed. The experiments were conducted on six speakers from a Soldier of Quarter Board corpus. Results showed that speaker-independent stress/neutral recognition rates are very high. One of the main drawbacks in most studies on emotion recognition is the lack of a benchmark database to test different algorithms. Also knowledge of ground truth regarding the presence of

stress or emotion is typically lacking. However in our case biometric measures confirm that subjects were indeed under stress. The proposed solution addresses how the stress/emotion classifier will behave when tested on open speakers not included in the training set.

7. ACKNOWLEDGEMENTS

We would like to thank our partners, James Meyerhoff, George Saviolakis, and Michael Koenig, affiliated with Dept. of Neuroendocrinology, Div. Neuroscience, Walter Reed Army Institute of Research (WRAIR), Silver Spring, Maryland for providing the SOQ speech corpus.

8. 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] J. H. L. Hansen, B.D. Womack, "Feature Analysis and Neural Network Based Classification of Speech Under Stress", *IEEE Trans. Speech Audio Process.*(4):307-313, 1996.
- [3] D. A. Cairns, J. H. L. Hansen, "Nonlinear Analysis and Detection of Speech under Stressed Conditions", *J. Acoust. Soc. Am.* (96)(6):3392-3400, 1994.
- [4] J.H.L. Hansen, "Analysis and Compensation of Speech under Stress and Noise for Environmental Robustness in Speech Recognition", *Speech Communication*, vol. 20(2), pp. 151-170, November 1996.
- [5] H. Teager, "Some Observations on Oral Air Flow During Phonation", *IEEE Trans. on Acoustics, Speech, and Signal Processing*, Vol.ASSP-28, No.5, pp. 599-601, Oct. 1990.
- [6] 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. 241-261, 1990.
- [7] J.F. Kaiser, "On a Simple Algorithm to Calculate the 'Energy' of a Signal", *ICASSP-90*, pp. 381-384, 1990.
- [8] J.F. Kaiser, "On Teager's Energy Algorithm, its Generalization to Continuous Signals," in *Proc. 4th IEEE Digital Signal Processing Workshop*, Sept 1990.
- [9] M.A.Oleshansky, and J.L. Meyerhoff, Acute catecholaminergic responses to mental and physical stressors in man. *Stress Medicine* 8:175-179, 1992.
- [10] M. Rahurkar, J.H.L. Hansen, M.A.Oleshansky, J.L. Meyerhoff, M. Koenig "Frequency Band Analysis for Stress Detection using a Teager Energy Operator Based Feature", *ICSLP-02*, Denver, Colorado.
- [11] J.H.L. Hansen, C. Swail, A.J. South, R.K. Moore, H. Steeneken, E.J. Cupples, T. Anderson, C.R.A. Vloeberghs, I. Trancoso, P. Verlinde, "The Impact of Speech Under 'Stress' on Military Speech Technology" published by NATO Research & Technology Organization RTO-TR-10, AC/323(IST)TP/5 IST/TG-01, March 2000 (ISBN: 92-837-1027-4).