# Robust Angry Speech Detection Employing a TEO-Based Discriminative Classifier Combination

*Wooil Kim and John H. L. Hansen*

Center for Robust Speech Systems (CRSS), Erik Jonsson School of Engineering & Computer Science
University of Texas at Dallas, Richardson, Texas, USA
{wikim,John.Hansen}@utdallas.edu, http://crss.utdallas.edu

## Abstract

This study proposes an effective angry speech detection approach employing the TEO-based feature extraction. Decorrelation processing is applied to the TEO-based feature to increase model training ability by decreasing the correlation between feature elements and vector size. Minimum classification error training is employed to increase the discrimination between the angry speech model and other stressed speech models. Combination with the conventional Mel frequency cepstral coefficients (MFCC) is also employed to leverage the effectiveness of MFCC to characterize the spectral envelope of speech signals. Experimental results over the SUSAS corpus demonstrate the proposed angry speech detection scheme is effective at increasing detection accuracy on an open-speaker and open-vocabulary task. An improvement of up to 7.78% in classification accuracy is obtained by combination of the proposed methods including decorrelation of TEO-based feature vector, discriminative training, and classifier combination.

**Index Terms**: angry speech detection, TEO-based feature, discriminative training, classifier combination.

## 1. Introduction

Reliable stress/emotion detection can be used to increase the performance of speech/speaker recognition systems for a range of applications including spoken dialog systems, cognitive task assessment, and spoken document retrieval. For this paper in particularly, angry speech detection can be effectively employed by industry for call center systems to improve costumer service. Recently, extensive research has been conducted in the speech area to improve the performance of the stress/emotion classification [1]-[6]. The TEO-based feature has been well known to be effective at representing traits of stressed state by reflecting variations in excitation characteristics included in speech signals [2][7]-[9].

In this study, a robust angry speech detection scheme is proposed, which is based on TEO-based feature extraction. Decorrelation processing is applied to the TEO-based feature vector to increase the ability of acoustic model training by decreasing the correlation between feature elements and vector dimensions. Minimum classification error training is employed to obtain a more discriminative angry speech model from other stressed styles such as loud and Lombard speech. In this study, no knowledge on phonetic information of the input speech is used, while past research for TEO-based feature processing focused mostly on vowels. A combination with conventional Mel frequency cepstral coefficients (MFCC) is considered to supplement performance of the TEO-based feature in unvoiced segments. Here, we present the effect of the combination of TEO-based feature and MFCC features on classification performance according to the combining domain (i.e., feature combination and classifier combination).

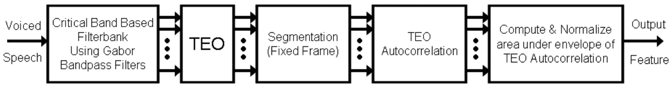
Voiced
Speech
Critical Band Based Filterbank Using Gabor Bandpass Filters
TEO
Segmentation (Fixed Frame)
TEO Autocorrelation
Compute & Normalize area under envelope of TEO Autocorrelation
Output
Feature

Figure 1: *TEO-CB-Auto-Env feature extraction flow [2].*

## 2. TEO-CB-Auto-Env: Critical Band Based TEO Autocorrelation Envelope

Fig. 1 shows a flow diagram of our previously proposed TEO-CB-Auto-Env feature extraction process [2]. The TEO [7][8] profile obtained from the Gabor bandpass filter output is segmented on a short-term basis, followed by an autocorrelation operation. The operation is intended to determine the level of "regularity" in the resulting segmented TEO response. Once the auto-correlation response is found, the area under the autocorrelation envelope is obtained and normalized. A single area coefficient is found which corresponds to each frequency band. The resulting vector of area coefficients has been shown to be large for neutral speech (i.e., speech has high "regularity") and low for speech that is produced with irregular excitation structure (i.e., for speech under stress and or speech under vocal fold pathology [9]). The TEO-CB-Auto-Env feature has been shown to reflect variations in excitation characteristics including pitch harmonics [2].

## 3. Minimum Classification Error Training

As a discriminative training method, minimum classification error (MCE) training [10] is employed in this study. The discriminant function for class $i$ can be defined as a log-likelihood function for each class model which is estimated as a Gaussian mixture model (GMM).

$$g_i(X) = \log\left[\sum_{k=1}^{K} \omega_{ik}\mathcal{N}(X;\boldsymbol{\mu}_{ik},\boldsymbol{\Sigma}_{ik})\right] \qquad (1)$$

This work was funded by Working Solutions Inc., and by University of Texas at Dallas under project EMMITT.

Using the defined discriminant function $g_i(X)$, the class misclassification measure can be defined as follows,

$$d_i(X) = -g_i(X) + \log\left[\frac{1}{M-1}\sum_{j,j\neq i}\exp\{g_j(X)\eta\}\right]^{1/\eta} \quad (2)$$

where $M$ denotes the class number and $\eta$ is a positive number controlling the relationship between competitive classes. In this study, we used 100 for $\eta$ in formulating the log-likelihood difference from the most competitive class as the misclassification measure. Eq.(3) shows the loss function used in this study,

$$\lambda_i(X) = \frac{1}{1+\exp(-\gamma d_i(X)+\theta)}, \quad (3)$$

where $\gamma$ controls the slope of the sigmoid function and $\theta$ is set to 0 in general.

In MCE training, the parameters of the discriminant function $g_i(X)$ are updated in a direction where the loss function $\lambda_i(X)$ is minimized. To minimize $\lambda_i(X)$, we maximize the misclassification measure $d_i(X)$, which improves the classifier discrimination across the different classes. As suggested in [10], the parameters of the discriminant function are updated as follows,

$$\begin{aligned}\widetilde{\mu}_{ikd}(n+1) &= \widetilde{\mu}_{ikd}(n) - \epsilon\frac{\partial\lambda_i(X)}{\partial\widetilde{\mu}_{ikd}}\\ \widetilde{\sigma}_{ikd}(n+1) &= \widetilde{\sigma}_{ikd}(n) - \epsilon\frac{\partial\lambda_i(X)}{\partial\widetilde{\sigma}_{ikd}}\end{aligned} \quad (4)$$

where,

$$\widetilde{\mu}_{ikd} = \frac{\mu_{ikd}}{\sigma_{ikd}}, \text{ and } \widetilde{\sigma}_{ikd} = \log(\sigma_{ikd}). \quad (5)$$

In Eq. (5), $\mu_{ikd}$ and $\sigma_{ikd}$ denote the $d$th component of the parameter vectors $\boldsymbol{\mu}_{ik}$ and $\boldsymbol{\Sigma}_{ik}$ for the model of class $i$. $\epsilon$ is a positive value ($< 1.0$) and set to 0.1 in our experiments.

## 4. Proposed Angry Speech Detection

Here, the proposed methods for robust angry speech detection are summarized as follows:

- Decorrelation of TEO-CB-Auto-Env feature parameter
- Discriminative training
- Feature combination
- Classifier combination

As presented in Sec. 2, the extracted TEO-CB-Auto-Env feature vector consists of the same element number as the number of Gabor bandpass filters employed for processing. In this study, we use 18 Gabor filters, resulting in an 18-dimensional TEO-based feature vector. Our previous study demonstrated that TEO-CB-Auto-Env feature values show a trend of frequency dependency (i.e., large values in the mid and small values for high frequency bands) [4], which suggests there exists a correlation relationship between feature elements of the TEO-based feature vector, since they are obtained from neighboring frequency bands. Here, we try to increase the model *trainability* by decreasing the vector size and correlation through employing a Discrete Cosine Transform (DCT) for the TEO-based feature extraction.

The TEO-CB-Auto-Env was originally designed to represent non-linear characteristics of the voiced sound production (e.g., vowels), showing effectiveness in stress/emotion detection [2][4]. However, sustained stress detection performance could depend on the ability of effective vowel sound detection. In this study, the entire speech duration is used without prior knowledge of phonetic information of the input speech. To sustain stress detection performance of the TEO-based feature in unvoiced sound segments, we believe that a conventional MFCC could be effective at increasing performance. In our recent study, a fusion of TEO and MFCC feature-based classifiers using the Adaboost algorithm demonstrated performance improvement for physical stressed speech detection using the UT-Scope database [11].

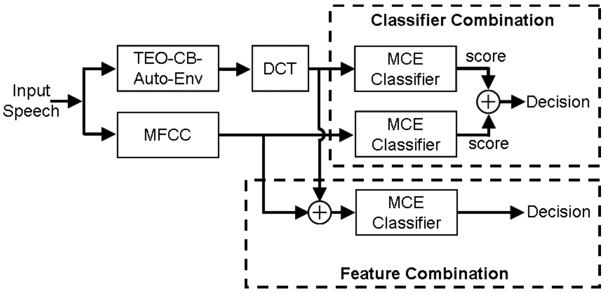

Figure 2: *Proposed feature and classifier combination scheme for anger speech detection.*

In this study, we employ two types of combination methods; (i) feature combination and (ii) classifier combination. In the feature combination approach, the MFCC feature vector is appended to the TEO-based feature vector. In the classifier combination approach, the classifier based on the TEO feature and the second classifier with an MFCC feature are composed at the decision stage by combining the likelihood scores from both classifiers with a set scale factor. We also will observe the performance trend for two combination approaches when applying the discriminative training method. It is expected that the classifier combination approach would be effective in taking an advantage of the improved classifier obtained by discriminative training under the restricted development data condition. Fig. 2 illustrates the proposed combination approaches for anger speech detection.

## 5. SUSAS Corpus

In this study, performance evaluation for anger speech detection was conducted using Speech Under Simulated and Actual Stress (SUSAS) [1][12]. The SUSAS corpus consists of five domains, encompassing a wide variety of stresses and emotions. A total of 32 speakers were employed to generate in excess of 16,000 isolated-word utterances. The five stress domains include: (i) talking styles (slow, fast, soft, loud, angry, clear, question), (ii) single tracking task or speech produced in noise (Lombard effect), (iii) dual tracking computer response task, (iv) actual subject motion-fear tasks (G-force, Lombard effect, noise, fear), and (v) psychiatric analysis data (speech under depression, fear, anxiety). The database consists of a common highly confusable vocabulary set of 35 aircraft communication words [12]. Simulated speech under stress data consists of the data from 10 stressed styles (talking styles, single tracking task and Lombard effect domains), while the actual speech under stress data consists of speech produced while performing either (i) dual-tracking workload computer tasks, or (ii) subject motion-fear tasks (subjects in roller-coaster rides).

Table 1: Data set for each evaluation session for open-speaker & open-vocabulary task.

| Data Set | Configuration | | Total Utterances |
|---|---|---|---|
| | Word† | Speaker‡ | (4 stresses) |
| Test | 7 (×2)* | 1 | 56 |
| Training | 21 (×2) | 8 | 1,344 |
| Development | 7 | 8 | 448 |

† No word overlap among test, training, and development sets

‡ No speaker overlap between test and training

* 2 times recorded

Table 2: Classification accuracy with MLE-training (%).

| Feature | Pairwise | 4-Class | Average |
|---|---|---|---|
| MFCC + logE (13) | 92.94 | 69.92 | 81.43 |
| TEO + logE (19) | 90.16 | 65.00 | 77.58 |
| DTEO + logE (13) | 90.96 | 66.75 | 78.85 |
| Feature Combination | Pairwise | 4-Class | Average |
| TEO + MFCC + logE (31) | 94.52 | 73.02 | 83.77 |
| DTEO + MFCC + logE (25) | 93.58 | 74.13 | 83.85 |

## 6. Experimental Results

A portion of the database from the SUSAS corpus was used for performance evaluation of the proposed methods in this study. We used four types of simulated stressed speech (neutral, angry, loud and Lombard) which were collected from 9 male subjects. Each stress set consists of 2 occurances of 35 words (i.e., total 70 words) making up 630 utterances, resulting in a total 2,520 utterance for all stresses. In order to have all utterances participate in the classification test, 5 different combinations of words were generated. Each combination consists of 7, 21, and 7 words entry for test, training, and development sets respectively, having no word overlap between the 3 sets. The 5 combinations of word sets were applied for all 9 speakers, conducting a total of 45 independent evaluation sessions. The data sets employed for each session are summarized in Table 1. Note that a speaker for an evaluation session does not overlap with speakers in training and development data sets. It is also noted that there is no word overlap among the test, training, and development sets at each session, formulating an *open-speaker* and *open-vocabulary* task.

For feature extraction, the TEO-CB-Auto-Env (TEO), the proposed DCT transformed TEO-CB-Auto-Env (DTEO), and the conventional Mel Frequency Cepstral Coefficients (MFCC) were employed. An analysis window of 32 msec duration is used with a 16 msec skip rate for 8-kHz speech data. For the MFCC, a standard algorithm suggested by the European Telecommunication Standards Institute (ETSI) was employed [13], where the 23-Mel-filterbank outputs are transformed to 12 cepstral coefficients adding log-energy (i.e., c1-c12, logE). For the TEO-based feature, 18 Gabor bandpass filters were used as suggested by our previous study [2].

The evaluation was conducted with two types of classification; (i) *pairwise* and (ii) *4-class* classifications. In the pairwise classification, neutral and angry speech are only used for a binary decision. For 4-class test, all 4 stresses (neutral, angry, loud, and Lombard) are used for a 4-class classification, which particularly evaluates the discrimination ability between angry speech and other classes such as loud and Lombard. Each model was estimated as a GMM which consists of 64 mixture components, which were obtained through a conventional maximum likelihood estimation (MLE) algorithm.

Table 3: Classification accuracy with MCE-training (%).

| Feature | Pairwise | 4-Class | Average |
|---|---|---|---|
| MFCC + logE (13) | 93.26 | 69.68 | 81.47 |
| TEO + logE (19) | 90.40 | 65.16 | 77.78 |
| DTEO + logE (13) | 91.11 | 67.30 | 79.21 |
| Feature Combination | Pairwise | 4-Class | Average |
| TEO + MFCC + logE (31) | 94.29 | 72.86 | 83.57 |
| DTEO + MFCC + logE (25) | 93.13 | 73.41 | 83.77 |

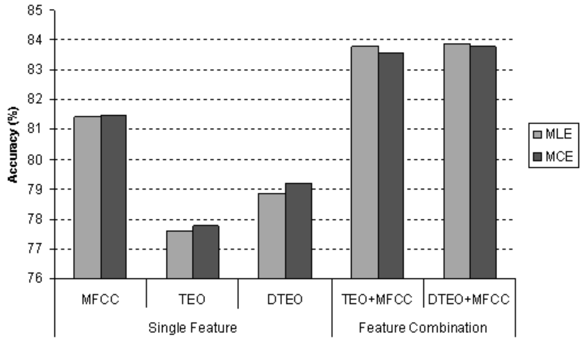

Figure 3: *Performance comparison of MLE and MCE training for different types of feature extraction.*

Table 2 shows performance of the feature extraction methods for angry speech detection, where each model was obtained by MLE training. The number with the feature type indicates the number of elements for each feature vector (i.e., vector dimension). Each value under “Pairwise” and “4-class” indicates the average value over the classification accuracies of neutral and angry speech for pairwise and 4-class tests respectively. It can be seen that the proposed DCT-TEO-based feature (DTEO) outperforms the conventional TEO-based feature (TEO). This result suggests that the employed decorrelation process for TEO-based feature was effective at improving the model ability to characterize the feature space, by reducing the correlation between feature elements and vector dimension. We can also see that a combination of the TEO features with MFCC feature vector produced performance improvement compared to all single feature cases. It is considered that the ability of the MFCC to characterize the spectral pattern of speech signals was effective at increasing angry speech detection.

Next, we conducted performance evaluation on the MCE-trained models. Mean and variance parameters of the acoustic models were updated through 7 iterations of MCE training over the development data presented in Table 1. Table 3 demonstrates that the employed MCE training brought consistently improved performance in the single feature cases compared to the results in Table 2. This suggests that MCE training was effective at reducing classification errors by generating more discriminative models between different stress classes. However, for the feature combination cases (i.e., TEO+MFCC+logE and DTEO+MFCC+logE), the MCE training was not as effective, even resulting in degraded performance in classification accuracy. We believe that the large number of feature dimension (31 and 25) could not effectively be applied to MCE training where the development data would not share many parts of the feature space with the test data. The development data consists of 448 utterances which do not overlap with test data in terms of speaker and vocabulary. We consider this a “sparse data” problem in MCE training for large dimensional feature vector,

Table 4: Classification accuracy of classifier combination with MLE-training (%).

| Classifier Combination | $\alpha$ | Pairwise | 4-Class | Average |
|---|---|---|---|---|
| TEO+logE & MFCC+logE | 0.25 | 93.97 | 73.89 | 83.93 |
| | **0.5** | **94.60** | **74.44** | **84.52** |
| | 0.75 | 92.78 | 72.54 | 82.66 |
| DTEO+logE & MFCC+logE | 0.25 | 93.58 | 73.33 | 83.45 |
| | **0.5** | **95.08** | **75.24** | **85.16** |
| | 0.75 | 93.33 | 74.21 | 83.77 |

Table 5: Classification accuracy of classifier combination with MCE-training (%).

| Classifier Combination | $\alpha$ | Pairwise | 4-Class | Average |
|---|---|---|---|---|
| TEO+logE & MFCC+logE | 0.25 | 94.53 | 74.05 | 84.29 |
| | **0.5** | **94.85** | **74.68** | **84.76** |
| | 0.75 | 93.10 | 72.86 | 82.98 |
| DTEO+logE & MFCC+logE | 0.25 | 94.21 | 73.33 | 83.77 |
| | **0.5** | **95.16** | **75.56** | **85.36** |
| | 0.75 | 93.58 | 74.21 | 83.89 |

and could be addressed by employing the model combination approach. Fig 3 summarizes the performance comparison of MLE and MCE training for each feature extraction method.

Table 4 shows the classification performance of the classifier combination approach as proposed in Sec. 4. The score scaling factor $\alpha$ and $(1 - \alpha)$ were applied to the likelihood scores of TEO-based and MFCC-based classifiers at the decision stage for the classifier combination. We obtained the best performance with 0.5 score factor for both TEO+MFCC and DTEO+MFCC cases. From the results, it can be seen that the classifier combination was effective at increasing classification accuracy compared to the feature combination approach shown in Table 2. Table 5 presents the performance of the classifier combination using the MCE-trained model. It is worth noting that the MCE-trained classifier combination method showed improved performance compared to MLE-based classifier combination method in both TEO+MFCC and DTEO+MFCC, while we obtained a performance degradation when applying MCE training to the feature combination for TEO+MFCC and DTEO-MFCC. These results suggest that classifier combination approach is effective at increasing the classification performance by independently utilizing the MCE-trained models under the restricted development data condition. Fig. 4 illustrates the final performance comparison of feature combination and classifier combination approaches employing MCE-training.

## 7. Conclusions

In this study, an effective angry speech detection scheme was proposed. Discrete Cosine Transform was applied to the TEO-based feature vector and minimum classification error training was employed for a discriminative training of the acoustic model for stressed speech. The combination with a conventional MFCC was applied in the feature domain and decision stage to integrate the reliability of the MFCC feature for representing spectral characteristics. Performance evaluation was conducted using the SUSAS corpus in a manner of open-speaker and open-vocabulary task. Experimental results demonstrated that the proposed method is considerably effective at increasing angry speech detection among 4 types of stressed speech; neutral, angry, loud and Lombard. We obtained up to 7.78% improvement in classification accuracy compared to the single TEO-based feature case, by combining several proposed methods including decorrelated TEO-based feature, discriminative training, and classifier combination approaches. Future work could consider potential speaker dependent traits for classifier combination factor $\alpha$.

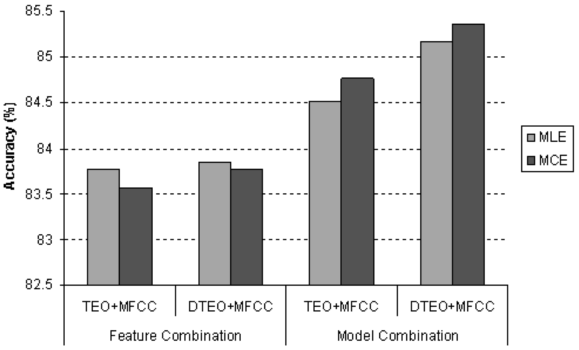

Figure 4: *Performance comparison of MLE and MCE training for feature and classifier combination approaches.*

## 8. References

[1] J.H.L. Hansen, "Analysis and Compensation of Speech Under Stress and Noise for Environmental Robustness in Speech Recognition," *Speech Communication*, 20 (2), pp.151-170, 1996.

[2] G. Zhou, J.H.L. Hansen, and J.F. Kaiser, "Nonlinear Feature-Based Classification of Speech Under Stress," *IEEE Trans. Speech & Audio Process.*, vol.9, no.3, pp.201-216, 2001.

[3] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. Taylor, "Emotion Recognition in Human-Computer Interaction," *IEEE Signal Processing Magazine*, vol.18, no.1., pp.32-80, 2001.

[4] M. Rahurkar, J.H.L. Hansen, J. Meyerhoff, G. Saviolakis, and M. Koenig, "Frequency Band Analysis for Stress Detection Using a Teager Energy Operator Based Feature," *ICSLP-2002*, pp.2021-2024, 2002.

[5] L. Devillers, and L. Vidrascu, "Real-life Emotions Detection with Lexical and Paralinguistic Cues on Human-Human Call Center Dialogs," *Interspeech2006*, 2006.

[6] V. Sethu, E. Ambikairajah and J. Epps, "Empirical Mode Decomposition based Weighted Frequency Feature for Speech-based Emotion Classification," *ICASSP2008*, pp.5017-5020, 2008.

[7] H. Teager, "Some Observations on Oral Air Flow During Phonation," *IEEE Trans. Acoustics, Speech, Signal Processing*, vol.28, no.5, pp.599-601, 1990.

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

[9] J.H.L. Hansen, L. Gavidia-Ceballos, and J.F. Kaiser, "A Nonlinear Operator-Based Speech Feature Analysis Method with Application to Vocal Fold Pathology Assessment," *IEEE Trans. Biomedical Engineering*, vol.45, no.3, pp.300-313, 1998.

[10] B. Juang, W. Chou, and C. Lee, "Minimum Classification Error Rate Methods for Speech Recognition," *IEEE Trans. Speech & Audio Process.*, vol.5, no.3, pp.257-265, 1997.

[11] S. Patil and J.H.L. Hansen, "Detection of Speech Under Physical Stress: Model Development, Sensor Selection, and Feature Fusion," *Interspeech2008*, pp.817-820, 2008.

[12] J.H.L. Hansen and S. Bou-Ghazale, "Getting Started with SUSAS: A Speech Under Simulated and Actual Stress Database," *Eurospeech'97*, pp.1743-1746, 1997.

[13] *ETSI standard document*, ETSI ES 201 108 v1.1.2 (2000-04), 2000.