Investigating the Robustness of Teager Energy Cepstrum Coefficients for Emotion Recognition in Noisy Conditions

Rui Sun, Elliot Moore II
Georgia Institute of Technology
210 Technology Circle
Savannah, Georgia 31407

Abstract

This paper investigated the robustness of Teager Energy Cepstrum Coefficient (TECC) in differentiating emotion categories for speech at different White Gaussian noise levels by comparing the performance with MFCC. Experiments involved the normalized squared error measurement, the multi-classes (four classes) emotion classification and the pair-wise emotion classification. This study included four emotion categories (neutral, happy, sad, and happy) from three databases (two English, one German). The result showed that TECC performed equally or outperformed MFCC in both multi-emotion and pair-wise emotion classifications at all noise levels for all three databases. Using TECC features only, up to 89% for the four-emotion classification and 99% for the pair-wise emotion classification accuracy rate could be achieved.

Introduction

Automated emotion detection is the attempt to quantify an abstract interpretation into objectively measured components of recorded human interaction. Emotion recognition in a noisy condition remains a challenging problem. The literature shows that Mel-Frequency Cepstrum Coefficients (MFCCs) exhibit robust performance in speech analysis in noisy environment, especially for speech recognition (Skowronski and Harris 2002; Dimitriadis, Maragos, and Potamianos 2005; Milner, Darch, and Vaseghi 2008; Muralishankar and O'Shaughnessy 2008; Zhang et al. 2009; Yu et al. 2008; Chu and Champagne 2008; Li and Huang 2011; Varela, San-Segundo, and Hernandez 2011; Milner and Darch 2011). On the other hand, the work on the use of Teager Energy Operator (TEO) (Caims and Hansen 1994; Zhou, Hansen, and Kaiser 2001; He et al. 2009; Sun and Moore 2011) in classifying emotion has shown that these features can provide valuable insight into distinguishing different types of emotional expression. These work motivate the study of emotion recognition using features combining the advantage of both MFCC and TEO, which has not been reported much yet. Teager Energy Cepstrum Coefficient (TECC) was first proposed by Dimitriadis and his colleges and studied to show its robust performance in speech recognition (Dimitriadis, Maragos, and Potamianos 2005; 2011). Motivated by the processing of MFCC, the extraction of TECC is similar to that of MFCC but using Teager Energy instead of the squared energy as the primary difference (details will be shown in the Section of Feature). In our study, the robustness of TECC in the application of emotion recognition for speech was investigated and compared with MFCC at different White Gaussian noise levels for three databases, two of which are English emotional databases and one is in German.

The paper is structured in the following way. The emotional speech data used in this study is described first. Then the extraction of features is explained, followed by the methodology of analysis. Finally, the results of analysis is presented and the conclusion is drawn.

Data

The emotional speech data in this study involves three databases, the Emotional Prosody Speech and Transcripts database (EPST), the Electromagnetic Articulography database (EMA), and the German Emotional Speech database (GES). To provide better comparability among databases, the four emotion categories (neutral, angry, sad, and happy) consistently presenting in all three databases are investigated. The number of sample units for each is shown in Table 1.

The EPST database (Liberman et al. 2002) contains recordings of emotional and semantically neutral speech spoken by seven native speakers (4 females and 3 males) of standard American English. All the speakers are professional actors. Each actor read short (4-syllables) dates and numbers (e.g. "five hundred one" or "august thirteenth") with the intent to express 15 different emotional categories ("neutral", "disgust", "panic", "anxiety", "hot anger", "cold anger", "despair", "sadness", "elation", "happy", "interest", "boredom", "shame", "pride" and "contempt") selected according to the Banse and Scherer's study (Banse and Scherer 1996). The speech was recorded at a sampling frequency of 22.05 kHz with 2-channel interleaved 16-bit PCM format (down sampled to 16 kHz for this study). The duration of each utterance varied approximately from 1 to 2 seconds. Speech data of emotion categories neutral, hot angry, sad, and happy from all seven actors in the EPST database were used in this study.

Table 1: Number of sample units for databases.

Emotions	neutral	angry	sad	happy	total
EPST	80	139	159	165	543
EMA	146	141	157	124	568
GES	79	127	62	71	339

The EMA database contains the emotional speech recordings of four emotions ("neutral", "angry", "sad", and "happy") from three English speakers (2 females an 1 male) (Lee, Kazemzadeh, and Narayanan 2005). A set of 14 sentences, which were mostly semantically neutral (e.g.,"Your grandmother is on the phone"), was uttered by each speaker to express four target emotions with five repetitions. The sentence duration varies from 1 to 4.5 seconds. The sampling rate of the recordings was 16kHz. Except for the missing sentences due to technical issues while recording, all four-emotion sentences from three speakers were included in this study.

GES is a German emotional database spoken by 10 speakers (5 females and 5 males) simulating seven emotions ("neutral", "disgust", "anxiety", "sad", "happy", and "boredom")(Burkhardt et al. 2005). The speech material was a set of 10 sentences with no semantically emotional bias covering everyday life content (e.g., "The cloth is lying on the fridge"). The sentence duration varies from 1 to 9 seconds. The speech was recorded at 48kHz and later down sampled to 16kHz. Speech sentences of neutral, angry, sad, and happy from all 10 actors in the GES database were used in this study.

Feature

The statistics of MFCC and TECC features and their derivatives (ΔMFCC and ΔTECC) were extracted and calculated to form the feature set in this study.

MFCCs were computed from the log-squared-energy in frequency bands distributed over a Mel-scale. The extraction of MFCC features for each speech sample was processed in five steps: (1) marked the voiced section of speech, (2) divided the voiced section into frames approximating four pitch periods in length with a 10ms step, (3) took the Fourier Transform on each frame, (4) mapped the power spectrums on to a Mel-scale, (5) took the log of the power at each Mel-scale band, (6) took the Discrete Cosine Transform (DCT) of Mel-log powers. The amplitude of the resulting spectrum was MFCC. The ΔMFCC feature was calculated by Eq. 1,

$$\Delta MFCC_j(i) = MFCC_j(i+1) - MFCC_j(i), \quad (1)$$

where $MFCC_j(i)$ is the j^{th} coefficient of MFCC from the i^{th} frame. The number of coefficients of MFCC used in this study is 12.

TECC was proposed with the motivation of the processing of MFCC feature and Teager Energy Operator (Dimitriadis, Maragos, and Potamianos 2005). Teager Energy was proposed by Teager based on his nonlinear model of the true source of sound production, which is actually the vortex-flow interactions (Teager 1980; Teager and Teager 1983). He developed the Teager Energy Operator supporting the observation that hearing is the process of detecting the energy. The TEO of discrete-time speech signal $x(n)$ can be calculated following Eq. 2 derived by Kaiser (Kaiser 1990),

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

where $x(n)$ is the n^{th} sample of signal. The extraction procedure of TECC was similar to MFCC but using Teager Energy $TEO[x(n)]$ instead of the squared energy $[x(n)]^2$ as the primary difference. The voiced speech was segmented into frame with length approximating four times of the pitch period with a step size of 10ms. The extraction on one frame of signal was demonstrated in Figure 1. The Gammatone filter is given by Eq. 3 in the time domain,

$$g(t) = At^{n-1}exp(-2\pi ERB(f_c)t)cos(2\pi f_c t), \quad (3)$$

where A, b, and n are Gammatone filter design parameters and f_c is the center frequency. According to (Irino and Patterson 1997; Dimitriadis, Maragos, and Potamianos 2005), the parameters are set as $b = 1.019$ and $n = 4$. Equivalent Rectangular Bandwidth (ERB) represents the bandwidth of filters, which is given by Eq. 4,

$$ERB(f) = 6.23(f/1000)^2 + 93.39(f/1000) + 28.52. \quad (4)$$

where f is the center frequency in Hz. And the filter gain A is set under the consideration that the frequency response at the center frequency equals to one. The filter placing is in Bark-scale (critical filterbank) (Zwicker and Terhardt 1980) and the number of filterbank in this study is 25. The ΔTECC feature is calculated by Eq. 5,

$$\Delta TECC_j(i) = TECC_j(i+1) - TECC_j(i), \quad (5)$$

where $TECC_j(i)$ is the j^{th} coefficient of TECC from the i_{th} frame. All features were quantified using seven statistics (i.e, the mean, median, minimum, maximum, standard deviation, range, and inter-quartile) across all frames of a sample to form the representation of an utterance. The feature extraction produced 168 MFCC features and 350 TECC features.

Methodology

The purpose of this study was to evaluate the robustness of the discrimination ability of TECC features in noisy conditions. Investigating the relationship between the robustness

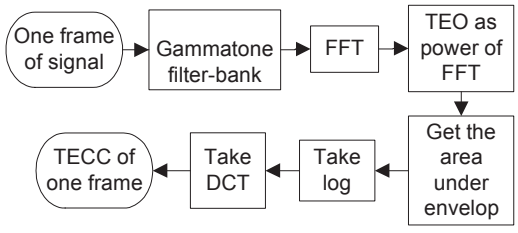

Figure 1: The flowchart of TECC extraction algorithm on one frame.

of features and the noise degree of speech requires emotional speech whose noise level is quantified and measurable. Therefore, five sets of data were created by adding White Gaussian noise to the "clean" speech dataset at five Signal Noise Ratio (SNR) levels from 20dB to 0dB with the step of 5dB. In total, six datasets (including the clean data) were available for each database (i.e.,five noisy and one clean).

It has been shown that acoustic features are speaker-dependent because they capture the characteristics of speakers (e.g., gender, language, culture) (Dromey, Silveira, and Sandor 2005). To eliminate the acoustic difference from factors other than emotion, speaker normalization was applied to the extracted features first. The processing of speaker normalization is shown in Eq. (6):

$$\hat{f_{i,j}} = \frac{f_{i,j} - mean(f_{i,j})}{std(f_{i,j})}, \tag{6}$$

where $f_{i,j}$ is the i^{th} feature descriptor for speaker j across the samples of all four emotions and std refers to the standard deviation. The normalization was conducted within database.

Given the normalized features, the normalized mean squared error (NSME) (Dimitriadis, Maragos, and Potamianos 2005; Yu et al. 2008) for MFCC and TECC was calculated at each noise level and compared. NSME is the measurement on the distance between feature of the noisy and clean speech from the same signal segment. The calculation of NMSE is shown in Eq. (7). It's defined as the average Euclidean distance between the "clean" and "noisy" features divided by the mean of "clean" feature vector norm (Dimitriadis, Maragos, and Potamianos 2005),

$$NMSE = \frac{mean(D(f_{i,clean}, f_{i,noisy})}{mean(|f_{i,clean}|)}, \tag{7}$$

where $D(f_{i,clean}, f_{i,noisy})$ is the Euclidean distance between the i^{th} feature in feature set of the clean speech and the noisy speech, and $|f_{i,clean}|$ is the vector norm of the i^{th} feature of the clean speech . The interpretation of NSME is that a smaller NSME value implies more robustness the feature possesses (i.e., NSME value is zero for the clean speech).

The robustness of the discrimination ability of features were evaluated in emotion recognition experiments. Using 5-fold cross-validation, the experiment built a four-class classifier on four subsets of data using a Support Vector Machine (SVM) and tested it on the other subset (using LibSVM tool (Chang and Lin 2011) in MATLAB, linear kernel). This procedure was repeated using another choice of training and testing sets till all sets has been tested. This classification was repeated 10 times for randomization. The further study was carried out as the discrimination ability in pair-wise emotion classification task. Four emotion categories formed six emotion pairs. One classifier was built for each pair using the liner kernel SVM with 5-fold cross-validation and the whole analysis was repeated 10 time as well.

Table 2: The normalized mean squared error (NSME) of MFCC/TECC at five SNR levels for three databases. The smaller value between MFCC and TECC under the same noisy condition using the same data is shown in **bold.**

SNR	EPST		EMA		GES	
	MFCC	TECC	MFCC	TECC	MFCC	TECC
0	0.58	**0.39**	0.52	**0.24**	0.56	**0.33**
5	0.50	**0.33**	0.44	**0.20**	0.47	**0.29**
10	0.41	**0.28**	0.35	**0.16**	0.38	**0.25**
15	0.32	**0.24**	0.27	**0.12**	0.30	**0.21**
20	0.24	**0.20**	0.20	**0.09**	0.22	**0.17**

Results

In this section, the normalized mean squared errors of MFCC and TECC on six datasets are reported. Then the classification results of multi-emotion and pair-wise emotion tasks are presented.

Mean Squared Error Analysis

Table 2 lists the NSME values at five SNR levels for three databases. It could be observed that, for all three databases, the values of MFCC and TECC decreases while the noise is reduced. It indicates the reliability of values in Table 2 according to the interpretation of NSME. It should be noticed that, at all noise levels of all three databases, the value of TECC is smaller than MFCC (value in bold). Especially for EMA, the all-level TECC values are less than half of MFCC. Based on the observation, the conclusion could be reached that TECC is more robust than MFCC facing additive noise in emotional speech. To further investigate the robustness of emotion-distinguishing ability of TECC, multi-emotion and pairwise-emotion classification experiments were conducted.

Emotion Recognition experiments

In the emotion recognition experiment, both TECC and MFCC features were applied in the multi-emotion classification (four-emotion) task first. The four-emotion classifier was built using LibSVM (Chang and Lin 2011) with the liner kernel. The accuracy rate (AR) was calculated as the average of those from 10 repetitions of classifications (5-fold cross-validation).

The accuracy rates at different noise levels are shown in Figure 2. From Figure 2 it's clear that the accuracy rates at all noise level using TECC are equal to or higher than MFCC for all three databases (i.e., AR using TECC is up to 71% in EPST, up to 89% in EMA, and up to 85% in GES for the four-emotion classification). When SNR equals to zero, TECC and MFCC performed equally. As the noise is reduced, using TECC improved the AR up to 38% for EPST, 9% for EMA, and 8% for GES. Overall, the ARs of EPST are relatively lower than EMA and GES. The possible explanation is that EPST contains 15 emotion categories while EMA has 4, and GES covers 7. The wider variety of emotion categories led to less acoustic difference between emotions

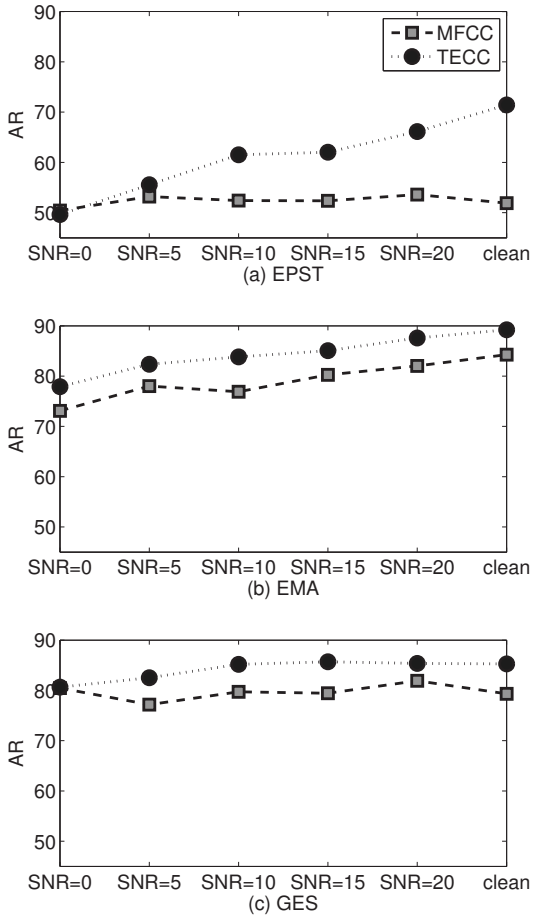

Figure 2: The accuracy rate (AR in %) of the 4-emotion classification using MFCC and TECC separately for three databases, (a)EPST, (b)EMA, and (c)GES.

for speech in EPST than the other two. Moreover, the robustness of emotion-distinguishing ability of TECC and MFCC is shown in the relationship between the variation of ARs with the change of noise levels.

For a better evaluation of the variation of ARs,the standard deviation of ARs at six noisy conditions using MFCC/TECC for each database is shown in Table 3. From Table 3, the standard deviation of MFCC and TECC is approximating equal. But for EPST, the variation of TECC is larger than MFCC. The reason for the larger variation of TECC is the increase in Figure 2(a). The conclusion could be reached that, both MFCC and TECC exhibit robustness in emotion recognition in noisy conditions while the overall AR of TECC is relatively higher. The larger variation of AR using TECC (in EPST) is caused by the performance improvement of ARs with the reduction of noise than MFCC.

The hypothesis is that the performance of TECC and

Table 3: The mean and standard deviation (std) of ARs over six noise levels for the 4-emotion classification using MFCC and TECC separately for three databases.

	EPST		**EMA**		**GES**	
	MFCC	TECC	MFCC	TECC	MFCC	TECC
mean	52.3	61.1	79.1	84.3	79.6	84.1
std	1.11	7.69	3.97	4.01	1.55	2.05

MFCC is not the same to all emotion categories. To test it, a pair-wise emotion classification experiment was conducted and the results are shown in Table 4. Since this experiment contains six emotion pairs at six noise levels for three databases. The resulting number of classification will be 108. Similar as the multi-classification task, for each classification, the accuracy rate was obtained as the average from 10 repetitions of 5-fold cross-validation classifications. Due to the large amount of resulting ARs, we show the mean AR across 6 noise levels, their standard deviation, and the "slope" of trend in Table 4 instead. The "slope" was calculated by averaging the ΔARs with the decrease of noise levels as shown in Eq. 8,

$$\Delta AR_{i,snr(n)} = AR_{i,snr(n+1)} - AR_{i,snr(n)}, \qquad (8)$$

where $AR_{snr(n)}$ is the AR at the n^{th} SNR level using feature i for each database, i.e., $\Delta AR_{(TECC,SNR=10)} = AR_{(TECC,SNR=20)} - AR_{(TECC,SNR=10)}$. A positive slope indicate the increase of AR with the reduction of noise.

From the mean ARs row of Table 4, the AR using TECC only reaches 74-85% for EPST, 86-99% for EMA, and 94-99% for GES in pair-wise emotion classification. From the standard deviation row of Table 4, the variation of ARs is quite small comparing with their mean values (up to 5% for EPST, 5% for EMA, and 4% for GES). This indicates the little effect from noise on the distinguishing ability of both TECC and MFCC. Comparing TECC with MFCC, the AR of TECC is increased by up to 15% for EPST (neutral-angry), 8% for EMA (angry-happy), and 5% for GES (angry-happy) than MFCC. What's more, in the "slope" rows, all slopes for TECC are positive, but some of MFCC is negative. It quantifies the observation of variation of MFCC in Figure 2, and emphasizes the reliability of the emotion-distinguishing ability of TECC on noise speech.

Overall, the ARs from EPST are lower than EMA and GES, which has been observed and explained in the multi-emotion task. Even though, the ARs for all three databases using TECC are fairly high, especially for EMA and GES (up tp 99%). Among six emotion pairs, the pair angry and happy possesses relatively lower ARs than other pairs for EMA and GES of both features. This observation could be explained by the conclusion that emotions with valence difference could be less captured by acoustics than arousal difference, which has been studied. This observation is not obvious in EPST data. The reason for this is that we chose "hot anger" in EPST, in which "cold anger" also uttered. Therefore, the angry in EPST was supposed to exhibit more difference in arousal than angry in other databases. As discussed with Table 4, the highest improvement of TECC than

Table 4: The mean, standard deviation (std), and the "slope" (slp) of accuracy rate from the pair-wise emotion (N:nutral, A:angry, S:sad, H:happy) classification of six noise levels using MFCC and TECC features for three databases.

		N-A		N-S		N-H		A-S		A-H		S-H	
		MFCC	TECC	MFCC	TECC	MFCC	TECC	MFCC	TECC	MFCC	TECC	MFCC	TECC
EPST	mean	69.9	80.3	76.8	84.4	79.4	86.1	67.5	74.2	72.0	77.5	71.7	74.0
	std	2.1	4.3	2.5	4.4	1.4	4.6	3.3	2.6	1.4	3.1	1.8	6.8
	slp	1.2	2.2	-0.9	2.3	-0.3	2.8	0.7	1.3	0.6	1.7	0.6	3.5
EMA	mean	96.6	98.6	85.6	88.1	96.0	97.5	97.9	98.5	79.6	85.9	94.3	97.9
	std	1.5	0.4	2.3	3.9	0.9	0.3	0.9	0.7	3.1	4.1	2.1	1.0
	slp	0.7	0.1	1.3	2.0	0.4	0.1	0.5	0.4	0.4	1.7	0.9	0.4
GES	mean	97.8	98.5	94.2	94.2	93.0	96.4	99.3	99.1	75.6	78.8	99.3	99.4
	std	0.8	0.2	1.1	1.4	2.9	1.0	0.3	0.2	2.1	3.0	0.5	0.2
	slp	-0.4	0.1	-0.1	0.4	-1.6	0.5	0	0.1	0.9	1.5	-0.2	0

MFCC happens in the neutral and angry pair for EMA and GES. This emphasized the performance of TECC when less acoustic difference exists.

Conclusion

This study investigates the robustness of TECC in emotion recognition facing additive noise at different levels. The results from three databases (two in English, one in German) highlight the robust emotion-discrimination ability of both TECC and MFCC. But the higher accuracy rate is achieved by TECC than MFCC. For the condition when SNR equals to zeros, TECC and MFCC performed similarly. While the noise level is reduced ($SNR = 5 \sim +\infty$), TECC outperformed MFCC in all emotion recognition tasks. Overall, using TECC features only, the up to 89% for the four-emotion classification and 99% for the pair-wise emotion classification accuracy rate could be achieved. Future work will involve the study using the real-life authentic emotional speech data in different speech quality conditions.

References

Banse, R., and Scherer, K. 1996. Acoustic profiles in vocal emotion expression. *Journal of ersonality and Social Psychology* 70:614–636.

Burkhardt, F.; Paeschke, A.; Rolfes, M.; Sendlmeier, W.; and Weiss, B. 2005. A database of german emotional speech. *in Proceedings of Interspeech, Lissabon* 1517–1520.

Caims, D., and Hansen, J. H. L. 1994. Nonlinear analysis and classification of speech under stressed conditions. *The Journal of the Acoustical Society of America* 96(6):3392–3399.

Chang, C.-C., and Lin, C.-J. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology* 2:27:1–27:27.

Chu, W., and Champagne, B. 2008. A noise-robust fft-based auditory spectrum with application in audio classification. *Audio, Speech, and Language Processing, IEEE Transactions on* 16(1):137 –150.

Dimitriadis, D.; Maragos, P.; and Potamianos, A. 2005. Auditory teager energy cepstrum coefficients for robust speech recognition. In *Interspeech'2005 - Eurospeech*.

Dimitriadis, D.; Maragos, P.; and Potamianos, A. 2011. On the effects of filterbank design and energy computation on robust speech recognition. *Audio, Speech, and Language Processing, IEEE Transactions on* 19(6):1504–1516.

Dromey, C.; Silveira, J.; and Sandor, P. 2005. Recognition of affective prosody by speakers of english as a first or foreign language. *Speech Communication* 47(3):351–359. doi: DOI: 10.1016/j.specom.2004.09.010.

He, L.; Lech, M.; Maddage, N.; and Allen, N. 2009. Emotion recognition in speech of parents of depressed adolescents. In *Bioinformatics and Biomedical Engineering , 2009. ICBBE 2009. 3rd International Conference on*, 1–4.

Irino, T., and Patterson, R. D. 1997. A time-domain, level-dependent auditory filter: The gammachirp. *The Journal of the Acoustical Society of America* 101:412–419.

Kaiser, J. F. 1990. On a simple algorithm to calculate the 'energy' of a signal. In *Acoustics, Speech, and Signal Processing, 1990. ICASSP-90., 1990 International Conference on*, 381–384 vol.1.

Lee, S.; Kazemzadeh, A.; and Narayanan, S. 2005. An articulatory study of emotional speech production. In *Eurospeech*, 497–500.

Li, Q., and Huang, Y. 2011. An auditory-based feature extraction algorithm for robust speaker identification under mismatched conditions. *Audio, Speech, and Language Processing, IEEE Transactions on* 19(6):1791 –1801.

Liberman, M.; Davis, K.; Grossman, M.; Martey, N.; and Bell, J. 2002. Emotional prosody speech and transcripts. *http://www.ldc.upenn.edu/ Catalog/ CatalogEntry.jsp? catalogId= LDC2002S28*.

Milner, B., and Darch, J. 2011. Robust acoustic speech feature prediction from noisy mel-frequency cepstral coefficients. *Audio, Speech, and Language Processing, IEEE Transactions on* 19(2):338 –347.

Milner, B.; Darch, J.; and Vaseghi, S. 2008. Applying noise compensation methods to robustly predict acoustic speech features from mfcc vectors in noise. In *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, 3945 –3948.

Muralishankar, R., and O'Shaughnessy, D. 2008. A comparative analysis of noise robust speech features extracted from

all-pass based warping with mfcc in a noisy phoneme recognition. In *Digital Telecommunications, 2008. ICDT '08. The Third International Conference on*, 180 –185.

Skowronski, M. D., and Harris, J. G. 2002. Increased mfcc filter bandwidth for noise-robust phoneme recognition. In *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*, volume 1, I–801 –I–804.

Sun, R., and Moore, E. I. 2011. Investigating glottal parameters and teager energy operators in emotion recognition. In D'Mello, S.; Graesser, A.; Schuller, B.; and Martin, J.-C., eds., *the 4th Affective Computing and Intelligent Interaction*, volume 6975 of *Lecture Notes in Computer Science*, 425–434. Memphis, TN: Springer.

Teager, H., and Teager, S. 1983. A phenomenological model for vowel production in the vocal tract. *Speech Science: Recent Advances* 73–109.

Teager, H. 1980. Some observations on oral air flow during phonation. *Acoustics, Speech and Signal Processing, IEEE Transactions on* 28(5):599–601.

Varela, O.; San-Segundo, R.; and Hernandez, L. 2011. Robust speech detection for noisy environments. *Aerospace and Electronic Systems Magazine, IEEE* 26(11):16 –23.

Yu, D.; Deng, L.; Wu, J.; Gong, Y.; and Acero, A. 2008. Improvements on mel-frequency cepstrum minimum-mean-square-error noise suppressor for robust speech recognition. In *Chinese Spoken Language Processing, 2008. ISCSLP '08. 6th International Symposium on*, 1 –4.

Zhang, J.; Li, G.-l.; Zheng, Y.-z.; and Liu, X.-y. 2009. A novel noise-robust speech recognition system based on adaptively enhanced bark wavelet mfcc. In *Fuzzy Systems and Knowledge Discovery, 2009. FSKD '09. Sixth International Conference on*, volume 4, 443 –447.

Zhou, G.; Hansen, J. H. L.; and Kaiser, J. F. 2001. Nonlinear feature based classification of speech under stress. *Speech and Audio Processing, IEEE Transactions on* 9(3):201–216.

Zwicker, E., and Terhardt, E. 1980. Analytical expressions for critical-band rate and critical bandwidth as a function of frequency. *The Journal of the Acoustical Society of America* 68(5):15231525.