# Data-driven Rescaled Teager Energy Cepstral Coefficients for Noise-robust Speech Recognition

Miau-Luan Hsu and Chia-Ping Chen
Department of Computer Science and Engineering
National Sun Yat-sen University
Kaohsiung, Taiwan
E-mail: M993040038@student.nsysu.edu.tw, cpchen@cse.nsysu.edu.tw

***Abstract*—We investigate data-driven rescaled Teager energy cepstral coefficients (DRTECC) features for noise-robust speech recognition. In the first stage, we apply a bank of auditory gammatone filters (GTF) and extract Teager-Kaiser energy (TE) estimates, which substitute the commonly used mel-spectrum. The output features of the first stage are called the Teager energy cepstral coefficients (TECC). In the second stage, we apply a piecewise rescaling operation of the cepstral coefficients of the zeroth order to bridge the difference between clean and noisy utterances. The segmentation point is determined by voice activity detection (VAD), and the proportional constants are data-driven. The resultant features are called DRTECC. The proposed features are evaluated on the Aurora 2.0 database. The relative improvements over the baseline MFCC features are significant.**

**Index Terms**: Teager energy estimation, noise-robust speech recognition, gamma-tone filters, energy rescale

## I. Introduction

Statistical methods have carried the technology of automatic speech recognition a long way for two decades [1]. However, when the training data and the test data are mismatched, the recognition accuracy still degrades very quickly [2]. Noises, including background noises and channel noises, are common sources for data-model mismatchedness. Unfortunately, noises are abundant in our daily lives, so the issue of noise-robustness must be properly addressed.

Many front-end methods have been proposed to treat the problem of mismatchedness, including speech enhancement techniques [3], [4], feature compensation methods [5], and noise-robust speech features [6], [7]. Speech enhancement is focused on improving the quality of noisy speech waveforms. Feature compensation aims to reduce the distortion of speech features via certain procedures beyond speech waveform.

Recently, new noise-robust features have been proposed based on auditory models and harmonic noise models (HNM) [8]. Specifically, the Teager energy estimates optimally approximate the speech energy. Furthermore, an auditory-inspired gammatone filter bank has been used for speech analysis. In this paper, the TECC features are based on the same ideas. We use an approximate implementation for the gammatone filter banks to improve computational efficiency.

It is well-known that energy is one of the most important speech features for recognition [9], [10], [11], [12], [13]. This has been verified by a series of experiments where we artificially replacing the noisy energy features by the clean energy features. In a noisy speech, the energy in a speech segment is generally higher than the energy in a non-speech segment. Therefore, it is beneficial to de-emphasize the low-energy segments and emphasize the high-energy segments. Many energy-rescaling methods have been proposed [14]. In this paper, we apply a piecewise rescaling operation to the zeroth-order cepstral coefficients, which is known to be highly correlated to speech energy.

The rest of paper is organized as follows. In Section II, we introduce the TECC features and our implementation. In Section III, we describe the data-driven approach used for obtaining rescaled TECC features. In Section IV, we present the evaluation results of the proposed methods. In Section V, we make concluding remarks and state future works.

## II. Feature Extraction

In Fig. 1, we depict the block diagrams for the extraction of the MFCC, TECC, RTECC, and DRTECC features. It can be seen that one of the major differences between TECC and MFCC is the set of filters used in the extraction. In traditional MFCC, triangular filters equally spaced in the mel-scale frequency axis are used, while in TECC, gamma-tone filters (GTF) are used. Another main difference with TECC is that the speech energy is estimated through Teager-Kaiser energy operator (TEO) on GTF output, instead of using DFT.

### A. Gamma-tone filter

A continuous-time gamma-tone filter is characterized by an impulse response function of

$$g(t) = at^{n-1}e^{-2\pi bt}\cos(2\pi ft + \phi), \tag{1}$$

where $a$ is the amplitude, $n$ is the order, $b$ is related to the filter bandwidth [15], $f$ is the central frequency, and $\phi$ is the phase shift. The discrete-time version of gamma-tone filter is implemented via a linear constant coefficient difference equation.

### B. Teager-Kaiser energy operator

TEO is a non-linear energy computation operator, which can increase the difference of energy between speech and noise. The continuous-time TEO applied to a real-valued signal $s(t)$ is given by

$$\psi\left[s(t)\right] = \left[\frac{d}{dt}s(t)\right]^2 - s(t)\left[\frac{d^2}{dt^2}s(t)\right] \tag{2}$$

The discrete-time representation is defined by

$$s_{teo}[n] = (s[n])^2 - s[n+1] \cdot s[n-1] \tag{3}$$

## III. Rescaling Features

Rescaling features for noise-robustness belongs to the category of feature compensation methods. The goal is to reduce the difference between the features of clean and noisy samples through rescaling operation on the raw features. The commonly used methods of histogram equalization and variance normalization are variants of rescaling features.

It is widely known that energy is one of the most important speech features for recognition. In the past, a rescaling scheme for the log energy feature has been proposed [14]. The basic idea is to emphasize the high-energy segments and de-emphasize the low-energy segments of speech. In order to retain the Teager energy operator to reduce the noise energy, we consider rescaling the zeroth cepstral coefficients $c_0$ in this paper. The rescaling function we investigate is a piecewise one. An utterance, say $u$, is processed independently of other utterances as follows.

- Find the maximum and minimum of the $c_0$ sequence of $u$. Denote the maximum by $M_u$ and the minimum by $m_u$.
- Consider frame $i$ with feature value $c_0[i]$. Let $r[i]$ be defined as
$$r[i] = \frac{c_0[i] - m_u}{M_u - m_u}.$$
Obviously, we have
$$0 \le r[i] \le 1.$$
- In rescaling, the rescaled feature is given by
$$\tilde{c}_0[i] = w[i]c_0[i], \tag{4}$$
where $w[i]$ is the weight for frame $i$. Assuming that the noise is quasi-stationary, higher energy segments are more likely to contain speech. Thus, it is desirable that
$$\begin{aligned} r[i] \approx 1 \quad &\longrightarrow \quad w[i] \approx 1, \\ r[i] \approx 0 \quad &\longrightarrow \quad w[i] \approx 0. \end{aligned} \tag{5}$$

There are many ways to implement the above idea shown in Eq. (5). We propose to use the following piecewise rescaling function

$$w[i] = \begin{cases} \left[\frac{\log(r[i]\times M)}{\log(M)}\right]^{\alpha_1}, & Y_{low}^{(i)} \le \theta \\ \left[\frac{\log(r[i]\times M)}{\log(M)}\right]^{\alpha_2}, & Y_{low}^{(i)} > \theta \end{cases} \tag{6}$$

A number of parameters are introduced in the implementation Eq. (6). $M$ is empirically set to 100, and $Y_{low}^{(i)}$ is low-frequency spectral magnitude [16]. Due to the proportion of speech and noise in low-frequency bands is constant, we set the range of frequency is from 0 to 50 Hz. $\theta$ is a threshold to determine the frame belongs to speech or non-speech, which is computed by

$$\theta = \frac{1}{P}\sum_{i=0}^{P-1} Y_{low}^{(i)} \tag{7}$$

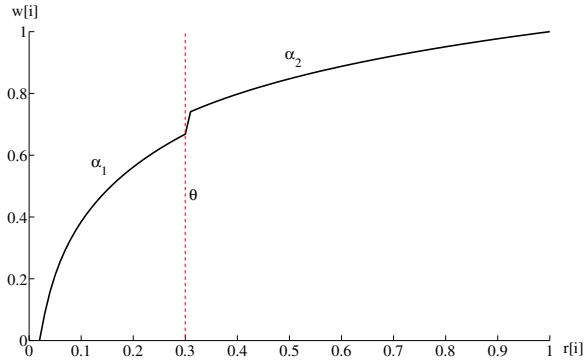


Fig. 2. Data-driven rescaling function. $\alpha_1$ and $\alpha_2$ are data-driven parameters.

where $P$ is assume to be the number of silence frames in the beginning of utterances. In this paper, $P$ is empirically set to 6. $\alpha_1$ and $\alpha_2$ are decided by minimizing the overall distortion on parallel training data sets. A grid search for the optimal parameters has been employed, with the following constraints

$$1 \le \alpha_2 < \alpha_1 \le 2.$$

The distortion is defined as

$$\begin{aligned} D(\alpha_1, \alpha_2, \theta) &= \sum_{u=1}^{U} \|C^u - N^u\| \\ &= \sum_{u=1}^{U} \left( \sum_{i=1}^{N_u} (C^u[i] - N^u[i])^2 \right)^{\frac{1}{2}} \end{aligned} \tag{8}$$

where $N^u$ is the rescaled feature sequence ($c_0$) of a noisy utterance, and $C^u$ is the rescaled feature sequence of the corresponding clean utterance. $C^u$ and $N^u$ are both computed by Eq. (4) and Eq. (6). $U$ is the total number of parallel utterances. We use the clean-train and multi-train data sets as parallel data to decide the parameters. The optimal parameters of TECC are

$$\alpha_1 = 1.3, \quad \alpha_2 = 1.0$$

which are then used in the rescaling of the test data as well. The features after the rescaling operation are called data-driven rescaled Teager-energy cepstral coefficients (DRTECC). A comparison between TECC, RTECC, and DRTECC feature sequences of a pair of parallel sample utterances has been presented in Fig. 3. It is clear that the rescaling operation is able to reduce the difference between clean and noisy utterances.

## IV. Experiments

### A. Recognition System Setup

We compare the rescaling features of MFCC, TECC, and AFE. The rescaled features are called DRMFCC, DRTECC, and DRAFE. The TECC-based feature vectors consist of the static features of 13 cepstral coefficients, i.e., $c_0, c_1, \ldots, c_{12}$. However, the log energy was used in MFCC and AFE. We use Eq. (8) to find the parameters of piecewise rescaling function

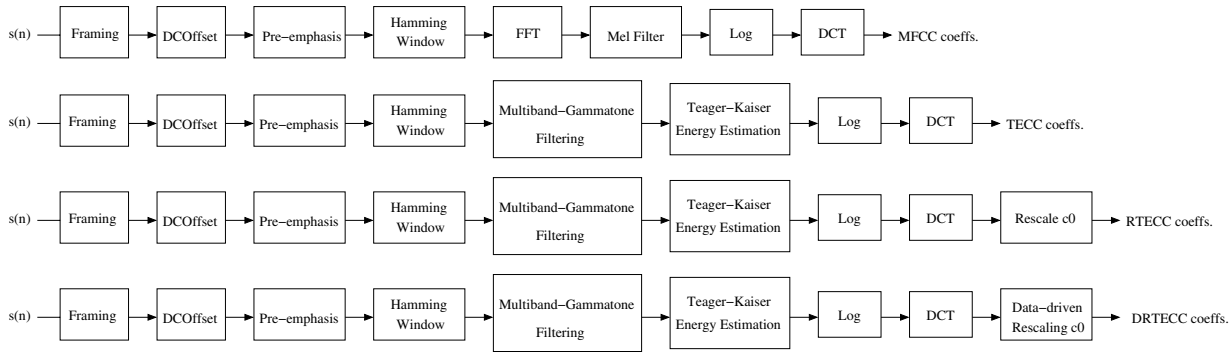


Fig. 1. Block diagrams for the extraction of the MFCC, TECC, RTECC, and DRTECC features.

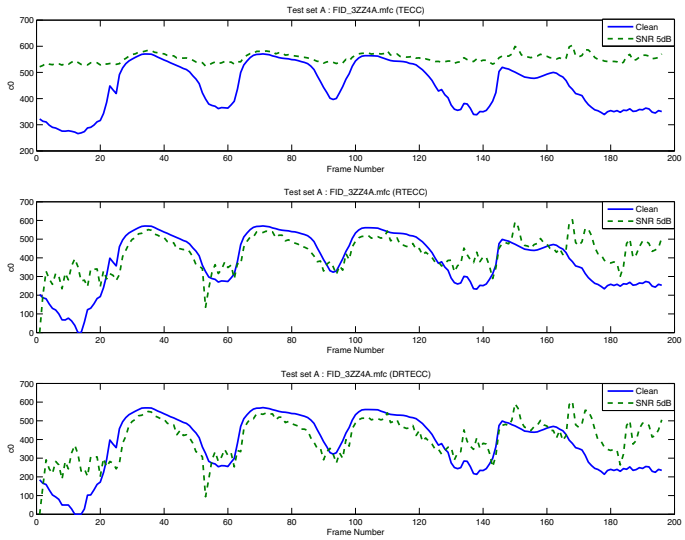


Fig. 3. Comparison of TECC, RTECC, and DRTECC $c_0$ feature sequences for a pair of parallel utterances, with ID FID_3ZZ4A.08.

for MFCC and AFE features. The optimal parameters for MFCC are

$$\alpha_1 = 1.3, \quad \alpha_2 = 1.0.$$

For the AFE features, the parameters were

$$\alpha_1 = 1.9, \quad \alpha_2 = 1.0.$$

During training and testing, velocity (delta) and acceleration (delta-delta) features are used.

The backend configuration in the standard Aurora evaluation framework [17] was used: 16-state wholeword hidden Markov models (HMM) for the digits, 3-state HMM for silence, and 1-state HMM for short pause. Each wholeword HMM state was associated with a 3-component Gaussian mixture model (GMM) for the state-emitting probability density. The silence/short-pause states used 6-component GMM. Furthermore, the short-pause HMM state was tied to the middle state of the silence HMM.

### B. Data

We used Aurora 2.0 speech database in the evaluation. Aurora 2.0 consists of continuous English digit utterances and it is widely used in noise-robust frontend evaluation. The utterances are artificially corrupted by 8 additive noises and 2 convolutional noises. According to signal-to-noise ratio (SNR) levels, the noisy utterances are ranged from 20 dB to -5 dB. The total number of training utterances is 8440, and the total number of test utterances for each noisy level is 1001.

### C. Results

Table I and Table II present the experiment results for the clean-train tasks and the multi-train tasks, with 0-20 dB SNR noisy test data, respectively. The rescaled log energy or $c_0$ features are called RMFCC, RTECC, and RAFE. The are called DRMFCC, DRTECC, and DRAFE if the parameters are data-driven. From Table I, we can see that the performance of DRMFCC, DRTECC, and DRAFE were consistently better than RMFCC, RTECC, and RAFE. In comparison of DRMFCC and DRTECC, we found that DRTECC outperformed

DRMFCC, in both clean-train and multi-train tasks. From Table II, the relative improvements of DRMFCC, DRTECC, and DRAFE are separately 10.36%, 16.11%, and 0.98%. The accuracy of DRTECC is much better than TECC and RTECC. The results with AFE were somewhat mixed. In the clean-train tasks, DRAFE was slightly better than RAFE, further justifying the usage of a data-driven approach. In the multi-train tasks, both RAFE and DRAFE were slightly better than AFE, meaning that rescaling operation did help to reduce errors.

TABLE I
RESULTS OF THE AURORA 2.0 CLEAN-TRAIN TASKS AVERAGED OVER TEST DATA WITH 0-20 DB SIGNAL-TO-NOISE RATIO.

| Feature | Set A | Set B | Set C | Avg. | Rel. imp. |
|---|---|---|---|---|---|
| MFCC | 61.34 | 55.75 | 66.14 | 60.06 | - |
| MFCC+MS | 66.18 | 70.81 | 64.88 | 67.77 | 19.30 |
| MFCC+MV | 70.18 | 70.77 | 66.37 | 69.65 | 24.01 |
| RMFCC | 74.60 | 74.51 | 65.23 | 72.69 | 31.62 |
| DRMFCC | 75.52 | 75.58 | 65.77 | 73.59 | 33.88 |
| TECC | 55.55 | 51.79 | 65.30 | 56.00 | - |
| TECC+MS | 66.92 | 71.52 | 67.67 | 68.91 | 29.34 |
| TECC+MV | 74.91 | 75.38 | 76.03 | 75.32 | 43.91 |
| RTECC | 77.36 | 77.40 | 66.97 | 75.30 | 43.86 |
| DRTECC | 79.09 | 80.19 | 72.15 | 78.15 | 50.34 |
| AFE | 86.69 | 85.57 | 82.81 | 85.47 | - |
| AFE+MS | 84.91 | 85.62 | 83.39 | 84.89 | -3.99 |
| AFE+MV | 76.80 | 76.85 | 74.39 | 76.34 | -62.84 |
| RAFE | 85.45 | 85.04 | 81.09 | 84.42 | -7.23 |
| DRAFE | 85.59 | 85.02 | 81.29 | 84.59 | -6.06 |

TABLE II
RESULTS OF THE AURORA 2.0 MULTI-TRAIN TASKS AVERAGED OVER TEST DATA WITH 0-20 DB SIGNAL-TO-NOISE RATIO.

| Feature | Set A | Set B | Set C | Avg. | Rel. imp. |
|---|---|---|---|---|---|
| MFCC | 87.82 | 86.27 | 83.78 | 86.39 | - |
| MFCC+MS | 88.72 | 87.79 | 87.25 | 88.06 | 12.27 |
| MFCC+MV | 89.67 | 88.07 | 86.10 | 88.32 | 14.18 |
| RMFCC | 89.36 | 86.54 | 85.51 | 87.46 | 7.86 |
| DRMFCC | 89.39 | 87.22 | 85.80 | 87.80 | 10.36 |
| TECC | 88.07 | 87.09 | 85.74 | 87.21 | - |
| TECC+MS | 89.14 | 89.33 | 89.66 | 89.32 | 16.50 |
| TECC+MV | 90.71 | 90.34 | 89.96 | 90.41 | 25.02 |
| RTECC | 90.04 | 87.61 | 86.38 | 88.33 | 8.76 |
| DRTECC | 90.02 | 89.10 | 88.10 | 89.27 | 16.11 |
| AFE | 91.79 | 90.76 | 89.11 | 90.84 | - |
| AFE+MS | 91.66 | 91.13 | 90.24 | 91.17 | 3.60 |
| AFE+MV | 90.99 | 89.68 | 88.11 | 89.89 | -10.37 |
| RAFE | 92.00 | 90.97 | 89.47 | 91.08 | 2.62 |
| DRAFE | 91.85 | 90.85 | 89.29 | 90.93 | 0.98 |

## V. CONCLUSIONS

In this research, we combined TECC-based front-end features and a data-driven rescaling method for noise-robust speech recognition. We used an energy-based voice activity detector and applied piece-wise rescaling function to the speech feature sequences of log energy or $c_0$. We employed the Aurora evaluation framework to evaluate the proposed method. Experiment results indicated that TECC improved over MFCC, and data-driven rescaling parameters outperformed fixed rescaling parameters.

## REFERENCES


[1] P. Krishnamoorthy and S. R. M. Prasanna, "Enhancement of noisy speech by temporal and spectral processing," *The 38th International Speech Communication Association (ISCA)*, vol. 53, no. 2, pp. 154–174, February 2011.

[2] C. Garreton, N. B. Yoma, and M. Torres, "Channel Robust Feature Transformation Based on Filter-Bank Energy Filtering," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 5, pp. 1082–1086, July 2010.

[3] D. Y. Zhao and W. B. Kleijn, "HMM-Based Gain Modeling for Enhancement of Speech in Noise," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 3, pp. 882–892, March 2007.

[4] J. Ming, R. Srinivasan, and D. Crookes, "A Corpus-Based Approach to Speech Enhancement From Nonstationary Noise," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 822–836, May 2011.

[5] K. Ngo, A. Spriet, M. Moonen, J. Wouters, and S. H. Jensen, "A combined multi-channel Wiener filter-based noise reduction and dynamic range compression in hearing aids," *Signal Processing*, vol. 92, no. 2, pp. 417–426, 2012.

[6] H. Veisi and H. Sameti, "The integration of principal component analysis and cepstral mean subtraction in parallel model combination for robust speech recognition," *The 17th International Conference on Digital Signal Processing (DSP 2011)*, vol. 21, no. 1, pp. 36–53, 2011.

[7] S. Furui, "Cepstral Analysis Technique for Automatic Speaker Verification," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 29, no. 2, pp. 254–272, 1981.

[8] D. Dimitriadis, P. Maragos, and A. Potamianos, "On the Effects of Filterbank Design and Energy Computation on Robust Speech Recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 6, pp. 1504–1516, 2011.

[9] T.-H. Hwang, "Energy contour extraction for in-car speech recognition," in *proceedings of 8th European Conference on Speech Communication and Technology (EUROSPEECH 2003)*, 2003.

[10] W. Zhu and D. O'Shaughnessy, "Log-energy dynamic range normalization for robust speech recognition," in *proceedings of 2005 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, 2005, pp. 245–249.

[11] T.-H. Hwang and S.-C. Chang, "Energy contour enhancement for noisy speech recognition," in *proceedings of 4th International Symposium on Chinese Spoken Language Processing (ISCSLP 2004)*, 2004, pp. 249–252.

[12] S. M. Ahadi, H. Sheikhzadeh, R. L. Brennan, and G. Freeman, "An energy normalization scheme for improved robustness in speech recognition," in *proceedings of 8th International Conference on Spoken Language Processing (ICSLP 2004)*, 2004.

[13] R. Chengalvarayan, "Robust energy normalization using speech/nonspeech discriminator for German connected digit recognition." in *proceedings of 6th European Conference on Speech Communication and Technology (EUROSPEECH 1999)*, 1999.

[14] H.-B. Chen, "On the study of energy-based speech feature normalization and application to voice activity detection," Master's thesis, National Taiwan Normal University, 2007.

[15] M. Slaney, "An Efficient Implementation of the Patterson-Holdsworth Auditory Filter Bank," *Apple Computer Perception Group Tech Rep*, no. 35, 1993.

[16] W.-H. Tu and J.-W. Hung, "Study on the Voice Activity Detection Techniques for Robust Speech Feature Extraction," in *Proceedings of 2007 Conference on Computational Linguistics and Speech Processing (ROCLING 2007)*, September 2007.

[17] D. Pearce and H. G. Hirsch, "The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *proceedings of 6th International Conference on Spoken Language Processing (ICSLP 2000)*, September 2000.