

Auditory Teager Energy Cepstrum Coefficients for Robust Speech Recognition

Dimitrios Dimitriadis¹, Petros Maragos¹ and Alexandros Potamianos²

¹National Technical University of Athens, School of ECE, Zografou, Athens 15773, Greece

²Technical University of Crete, Dept. of ECE, Chania 73100, Greece

Email: [ddim,maragos]@cs.ntua.gr, potam@telecom.tuc.gr

Abstract

In this paper, a feature extraction algorithm for robust speech recognition is introduced. The feature extraction algorithm is motivated by the human auditory processing and the nonlinear Teager-Kaiser energy operator that estimates the true energy of the source of a resonance. The proposed features are labeled as Teager Energy Cepstrum Coefficients (TECCs). TECCs are computed by first filtering the speech signal through a dense non constant-Q Gammatone filterbank and then by estimating the “true” energy of the signal’s source, i.e., the short-time average of the output of the Teager-Kaiser energy operator. Error analysis and speech recognition experiments show that the TECCs and the mel frequency cepstrum coefficients (MFCCs) perform similarly for clean recording conditions; while the TECCs perform significantly better than the MFCCs for noisy recognition tasks. Specifically, relative word error rate improvement of 60% over the MFCC baseline is shown for the Aurora-3 database for the high-mismatch condition. Absolute error rate improvement ranging from 5% to 20% is shown for a phone recognition task in (various types of additive) noise.

1. Introduction

Despite recent advances in the state-of-the art of automatic speech recognition (ASR), speech recognition in noisy conditions remains an open research problem. Robust speech recognition, in general, is an important research area; performance of spoken dialogue systems deployed in the field often degrades due to adverse and (often) unexpected environmental conditions. The techniques that have been proposed in the literature for improving the robustness of speech recognition in noise mainly fall into three categories: acoustic model adaptation algorithms, speech enhancement algorithms and robust feature extraction algorithms. In this paper, we concentrate in the problem of robust feature extraction. The Teager energy feature set is proposed motivated by speech perception considerations and the nonlinear Teager-Kaiser operator that estimates the energy of the source of a resonance signal.

The most widely used speech recognition features are the Mel Frequency Cepstrum Coefficients (MFCCs). MFCCs are computed from the log-energies in frequency bands distributed over a mel scale. The wide-spread use of the MFCCs is due to the low complexity of the estimation algorithm and their good performance for ASR tasks under clean matched conditions [1]. However, MFCCs are easily affected by common frequency-localized random perturbations, to which human perception is largely insensitive. MFCC performance degrades rapidly in the presence of noise and performance degradation is directly proportional to the signals’ SNR. MFCC’s lack of robustness in noisy or mismatched conditions have led many researchers to investigate robust variants of MFCCs or novel feature extrac-

tion algorithm altogether. Much of these research is motivated by models of human perception, e.g., the RASTA [4] and PLP features [3]. In this paper, we design a robust front-end that is motivated from auditory perception and uses a dense (in frequency) bank of Gammatone filters. The filter bandwidths are proportional to the auditory Equivalent Rectangular Bandwidth (ERB) function as described in [6, 7, 8].

The short-time average of the signal squared is widely used as an ad hoc approximation of the energy of the signal’s source. For resonance signals, the *Teager-Kaiser Energy* and the nonlinear energy operator Ψ provide a good estimation of the “real” source energy. Recently, Teager energy has been used for speech recognition in [10, 12]. In this paper, we extend this work and design a front-end that combines an auditory-motivated filterbank with the Teager energy estimation method. The proposed features labeled *auditory Teager Energy Cepstrum Coefficients* (TECCs) are evaluated on speech recognition tasks in noise and are shown to be more robust than the MFCCs. Robustness is shown both in the mean square error sense and in terms of speech recognition accuracy.

The organization of this paper is as follows: in Section 2, we provide the theoretical background of the nonlinear *Teager-Kaiser energy operator* and the *Human Auditory filterbank*. In Section 3, the proposed feature extraction algorithm is presented. In Section 4, the performance of the proposed features is evaluated under noisy recording conditions both in terms of mean squared error and in terms of recognition performance. Conclusions are presented in Section 5.

2. Theoretical Background

2.1. Teager-Kaiser Energy Operator

Newton’s law of motion for an oscillator with mass m and spring constant k states that

$$\frac{d^2x}{dt^2} + \frac{k}{m}x = 0$$

and its solution consists of a signal $x(t) = a \cos(\phi(t))$. The system’s total energy E is the sum of the kinetic and potential energy and is given by

$$E = \frac{1}{2}kx^2 + \frac{1}{2}m\dot{x}^2 \Rightarrow E = \frac{1}{2}m\omega^2a^2 \quad (1)$$

where $\omega = d\phi(t)/dt$.

Taking this analysis under consideration, Teager and then Kaiser [9], proposed the *Teager-Kaiser Operator* Ψ

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

When applied to an AM-FM signal $x(t) = a(t) \cos(\phi(t))$, the Ψ operator yields

$$\Psi[x(t)] \cong a^2(t)\dot{\phi}^2(t) \quad (3)$$

Herein, instead of using the “traditional” signal energy approximation of x^2 (that only takes into account the kinetic energy of the signal’s source) we will use the Teager-Kaiser energy operator for computing the “true” source energy. The short-time average of the output of the energy operator will be used for feature estimation. The Teager-Kaiser estimated energy incorporates both amplitude and frequency information; the hope is that the additional information in the estimated energy can be translated into improvement in speech recognition accuracy.

2.2. Auditory Filterbank

Human auditory processing relies on a set of dense (in frequency) asymmetrical filters that estimate the activity in each frequency band. The notion of the Equivalent Rectangular Bandwidth (ERB) can be used to quantify the bandwidth of asymmetrical filters like the auditory ones. Specifically, given the magnitude of a filter’s frequency response $|H(f)|$ and the filter’s maximum gain $|H(f_{max})|$ at frequency f_{max} the filter’s ERB (in Hz) is defined as

$$ERB = \frac{\int |H(f)|^2 df}{|H(f_{max})|^2} \quad (4)$$

The ERB is the equivalent bandwidth of an orthogonal filter with constant gain $|H(f_{max})|$ and energy equal to the original filter’s energy (the filter’s energy is defined as the integral of the filter’s frequency response squared).

Recent studies [6, 7, 8] have shown that the human physiology dictates that the auditory filter bandwidths are given by the $ERB(f)$ function

$$ERB(f) = 6.23(f/1000)^2 + 93.39(f/1000) + 28.52 \quad (5)$$

where f is the filter center frequency in Hz. Moreover, the filter placing is equidistant in the *critical* (bark) frequency scale

$$bark(f) = \frac{26.81f}{f + 3920} - 0.53 \quad (6)$$

where $0 \leq f \leq F_s/2$ and F_s is the sampling frequency of the signal. A good approximation of the auditory filters are the asymmetrical Gammatone filters with impulse response

$$g(t) = At^{n-1} \exp(-2\pi b ERB(f_c)t) \cos(2\pi f_c t) \quad (7)$$

where A , b , n are the Gammatone filter design parameters and f_c is the center frequency of the filter. In [8], it is proposed that the auditory filters should have $b = 1.019$ and $n = 4$. Thus, the filter frequency response $G(\omega)$ is given by

$$G(\omega) = \frac{A}{2} \frac{6}{(2\pi b ERB(f_c) + j(\omega - \omega_c))^4} + \frac{A}{2} \frac{6}{(2\pi b ERB(f_c) + j(\omega + \omega_c))^4} \quad (8)$$

Moreover, the filter gain A is set taking under consideration that $|H(\omega_c)| = 1$ and is equal to

$$A = \frac{1}{\sum_{k=1}^N t^{n-1} \exp(-2\pi b ERB(f_c)t)} \quad (9)$$

where N is the length of the discretized impulse response in samples.

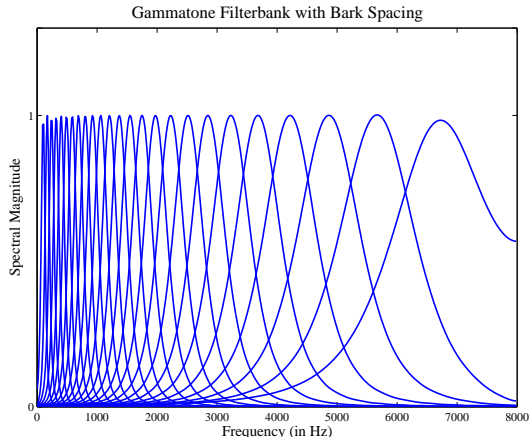


Figure 1: A Gammatone filterbank with 25 filters and bandwidths $1.5 ERB(f)$.

The auditory filterbank proposed above is not constant-Q¹ and emphasizes the lower part of the frequencies where the main part of the acoustic information is located. Mel-spaced filterbanks used for MFCC feature extraction in speech recognition tasks [1] use symmetric filters and constant-Q filterbanks. The main differences between the proposed filterbank and the typical one used for MFCC estimation are the type of filters used and their corresponding bandwidths. As we will show in the next section, constant-Q filterbanks are not always the best choice for speech recognition tasks; mimicking the human auditory system can provide superior results, especially robustness in noise or adverse recording conditions.

The Gammatone filterbank presented above, with filters placed according to the bark scale and with bandwidths given by the $ERB(f)$ is a good approximation of the human auditory system [3, 4, 6]. The human ear employs several thousand filters and the corresponding filterbank is very dense (in frequency). In this paper, we experiment with two parameters to create a family of Gammatone filterbanks:

- The number of filters (# Filter) in the filterbank, i.e., the filterbank density.
- The bandwidth of the filters as a percent of $ERB(f)$; the bandwidth of the filter at center frequency f is obtained by multiplying the filter bandwidth curve $ERB(f)$ by the parameter F .

Experimental results provided in the next section show that both parameters are important for robust speech recognition. The range of parameters we have experimented is 20 - 40 for the number of filters and 1.0 - 2.0 for the bandwidth multiplying factor F . An example of the Gammatone filterbank employing 25 filters and with $F = 1.5$ is shown in Fig. 1.

3. Auditory Teager Energy Cepstrum Coefficients

The *auditory Teager Energy Cepstrum Coefficients* (TECCs) are extracted from the speech signal according to the following steps:

¹Constant-Q filterbanks the ratio of frequency spacing over bandwidth between two neighbouring filters constant.

- (i) Use the bi-parametric family of Gammatone filterbanks defined in Eqs. (7), (8) with parameters the number of filters (20-40) and the bandwidth multiplying factor F (1.0-2.0) to bandpass the speech signal. The filter spacing is linear in the bark scale.
- (ii) Estimate the logarithm of the short-time average of the Teager-Kaiser energy operator for each one of the band-passed signals. The short-time averaging window duration and window shift are the same as for the “standard” MFCC front-end (30 and 10 msec respectively).
- (iii) Estimate the cepstrum coefficients of the short-time average Teager energy using the discrete cosine transform (DCT), and
- (iv) Truncate the cepstrum coefficients to keep the first 13 coefficients (including the zeroth coefficient C_0) similarly to the “standard” MFCC front-end.

The first two steps are the main differences between TECC and MFCC feature extraction, namely the auditory filterbank and the short-time Teager energy computation. The “standard” MFCC front-end uses filters with frequency response that is triangular in shape and constant-Q (50% filter frequency response overlap). The proposed auditory TECCs use filters that are smoother and broader [11] than the MFCC triangular filterbank (the bandwidth of the filter is controlled by the *ERB*-curve and the bandwidth multiplication factor F). Also, the TECC filterbank is more dense in frequency (controlled by the number of filters parameter) and spaced according to the bark-scale rather than the mel-scale. By experimenting with filter frequency response, density and bandwidth, and by using the more informative Teager energy estimate, significant improvements in recognition accuracy are shown in the next section (TECCs vs. MFCCs).

4. Experiments and Results

In this section, we investigate the robustness of TECCs in noise by artificially injecting various types of noise to the speech signal and computing the normalized mean squared error (for both MFCCs and TECCs). We then present speech recognition experiments in noisy recording conditions. Results from both a connected-digit recognition task (Aurora-3) and a phone recognition task (TIMIT+Noise) are presented.

4.1. Mean Square Error Analysis

For the needs of this experiment, we have created a ‘TIMIT+Noise’ database by artificially adding babble, white, pink or car noise (noise samples are from the NOISEX database) at 10 dB to the test set of the TIMIT database. Feature robustness is computed in terms of normalized mean squared error (NMSE). The NMSE is defined as the average Euclidean distance between the “clean” and “noisy” features divided by the mean “clean” feature vector norm. The feature vector used in the Euclidean distance computation consists of 12 MFCCs or TECCs excluding the zeroth cepstrum coefficient C_0 . “Clean” and “noisy” features refer to the features computed on the same speech segment before and after the addition of noise.

As shown in Table 1, the proposed TECC features have smaller NMSE than the MFCC features for all noise types. The relative improvement in NMSE ranges between 20% and 30% for various type of noise. TECC feature robustness is affected by the choice of bandwidth multiplication factor F and the number of filters in the auditory filterbank; on average

	Babble	White	Pink	Car
MFCC (baseline)	0.523	0.646	0.612	0.464
TECC ($F=1.5, \#Flt=25$)	0.393	0.491	0.449	0.302
TECC ($F=2.0, \#Flt=25$)	0.408	0.489	0.460	0.322
TECC ($F=1.5, \#Flt=30$)	0.391	0.463	0.435	0.322
TECC ($F=2.0, \#Flt=30$)	0.386	0.483	0.441	0.323

Table 1: Normalized mean squared error of MFCCs and TECC for Various Types of Additive Noise (at SNR = 10dB).

(across noise types) best results are obtained for F around 1.5 and approximately 30 filters. Based on the NMSE analysis results shown in Table 1 we conclude that TECCs are significantly more robust than MFCCs in the presence of additive noise; this claim is also supported by the speech recognition experiments discussed in the next Section.

4.2. Recognition Experiments

We have applied the proposed TECC features to the Aurora-3 speech database (Spanish) connected-digit recognition task and to the TIMIT+Noise database (see previous section) phone recognition task. The Aurora-3 database contains recordings, sampled at 8 kHz, from 2 different microphones, at 3 driving conditions. These recordings are mixed to create 3 different training/testing scenarios, the *Well-Matched* (WM) scenario, the *Medium-Mismatch* (MM) scenario, where the mismatch is mainly due to the usage of different microphones and the *High-Mismatch* (HM) scenario with different noise levels in the training and the testing sets. The ASR experiments have been performed using the HMM-based HTK Toolkit [13]. Context-independent, 14-state, left-right word HMMs with 16 Gaussian mixtures per state are used for the Aurora-3 task. For the TIMIT+Noise task, the models used are 3-state, left-right phone HMMs with 16 Gaussian mixtures per state. The grammar used for both cases is an all-pair, unweighted grammar (open-loop). For the Aurora-3 task three models are trained, one for each of the WM, MM and HM recording conditions. For the TIMIT+Noise task, the HMM models are trained under clean recording conditions and tested in the noise-corrupted (at 10 dB SNR) versions of the test set, i.e., there is mismatch in the training and test conditions.

The feature vector consists of 39 coefficients for both the MFCC and TECC features, i.e., the zeroth cepstrum coefficient plus the first 12 cepstrum coefficients and their 1st and 2nd time-derivatives. Cepstral Mean Subtraction (CMS) is applied to the proposed features (both during training and testing). The analysis window duration is 30 msec and the window update is 10 msec.

In Tables 2 and 3, the recognition results are presented for the Aurora-3 and the TIMIT+Noise tasks, respectively. For the Aurora-3 task, the TECC features are shown to significantly outperform the MFCCs for the High-Mismatch (HM) condition; there is no significant difference between TECCs and MFCCs for the WM and MM conditions. For the TIMIT+Noise task, the TECCs significantly outperform the MFCCs for all noisy conditions; for clean conditions the MFCCs slightly outperform the TECCs. The best TECC results are obtained for an auditory filterbank with 30 filters and multiplicative bandwidth factor of $F = 1.5$. Overall, the TECCs outperforms the MFCCs under all additive noise conditions for both the Aurora-3 and TIMIT+Noise tasks. Relative error rate improvements range from 5% to 60% depending on the type and level of noise.

Aurora-3 Database, Spanish Task						
Features	Scenario	WM	MM	HM	Average	Aver. Rel. Improv. (%)
Aurora Frontend (WI007)		92.94	80.31	51.55	74.93	-
MFCC (Baseline with CMS)		93.68	92.73	65.18	83.86	35.62
TECC† (F=1.5, # Filter=25)		94.33	91.29	86.31	90.64	62.66
TECC† (F=2.0, # Filter=25)		93.92	90.42	83.82	89.39	57.68
TECC† (F=1.5, # Filter=30)		93.93	91.80	86.85	90.86	63.54
TECC† (F=2.0, # Filter=30)		93.32	90.92	84.22	89.49	58.08
†TECC + C0 + 1 st +2 nd Time Derivatives + CMS						

Table 2: Word Accuracies (%) for the MFCC and TECC features for the Aurora-3 Spanish task.

TIMIT+Noise Tasks (for SNR=10 dB)						
	TIMIT	TIMIT +Babble	TIMIT +White	TIMIT +Pink	TIMIT +Car	Aver. Rel. Improv. (%)
MFCC (Baseline with CMS)	58.40	27.71	17.72	18.60	52.75	-
TECC† (F=1.5, # Filter=25)	55.71	39.55	33.54	37.00	50.82	23.66
TECC† (F=2.0, # Filter=25)	57.40	38.35	33.17	36.32	48.20	21.84
TECC† (F=1.5, # Filter=30)	57.15	39.72	33.97	37.56	50.10	24.73
TECC† (F=2.0, # Filter=30)	57.86	37.81	32.72	36.18	47.61	21.12
†TECC + C0 + 1 st +2 nd Time Derivatives + CMS						

Table 3: Phone Accuracies (%) for the MFCC and TECC features for the TIMIT+Noise task.

5. Conclusions

The proposed TECC features have been shown to be more robust than MFCCs in additive noise. TECC feature robustness was demonstrated both in terms of mean square error analysis and speech recognition performance in two tasks (Aurora-3, TIMIT+Noise). For the TIMIT+Noise phone recognition task the average relative improvement for various types of noise was 24% at 10 dB SNR. Up to 60% relative error rate reduction was achieved for the High-Mismatch Aurora-3 task over the baseline MFCC performance. For clean conditions and convolutional noise the TECCs performed similarly to the MFCCs. The increased robustness of the TECCs could be due both to the auditory filterbank design and the Teager energy estimation, however, from preliminary experiments it appears that a large portion of the improvement is due to the use of the auditory filterbank. Research is under way to quantitatively investigate the source of the increased TECC robustness in additive noise.

Acknowledgments: This research work was partially supported by the IST EU FP6 research programs HIWIRE and MUSCLE and by the NTUA basic research program ‘Protogoras’.

6. References

- [1] S. B. Davis and P. Mermelstein, “Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences”, *IEEE Trans. ASSP*, vol. 28, No. 4, pp. 357–366, Aug. 1980.
- [2] D. Dimitriadis and P. Maragos, “Robust Energy Demodulation Based on Continuous Models with Application to Speech Recognition”, in *Proc. of Eurospeech-03*, Geneva, Sept. 2003.
- [3] H. Hermansky, “Perceptual Linear Predictive (PLP) Analysis of Speech”, *J. Acoust. Soc. Am.*, Vol. 87, No. 4, pp. 1738–1752, 1990.
- [4] H. Hermansky and N. Morgan, “RASTA Processing of Speech”, *IEEE Trans. SAP*, Vol. 2, No. 4, pp. 578–589, 1994.
- [5] F. Jabloun, A. E. Cetin, and E. Erzin, “Teager Energy Based Feature Parameters for Speech Recognition in Car Noise”, *IEEE SPL*, Vol. 6, No. 10, pp. 259–261, Oct. 1999.
- [6] O. Ghizta, “Auditory Models and Human Performance in Tasks Related to Speech Coding and Speech Recognition”, *IEEE Trans. SAP*, Vol. 2, No. 1, pp. 115–132, 1994.
- [7] B. R. Glasberg and B. C. J. Moore, “Derivation of Auditory Filter Shapes from Notched-Noise Data”, *Hear. Res.*, Vol. 47, pp. 103–138, 1990.
- [8] T. Irino and R. D. Patterson, “A Time-Domain, Level-Dependent Auditory Filter: The Gammachirp”, *J. Acoust. Soc. Am.*, Vol. 101, pp. 412–419, 1997.
- [9] J. F. Kaiser, “On A Simple Algorithm to Calculate the ‘Energy’ of a Signal”, in *Proc. ICASSP-90*, Albuquerque, New Mexico, pp. 381–384, April 1990.
- [10] A. Potamianos and P. Maragos, “Time-Frequency Distributions for Automatic Speech Recognition”, *IEEE Trans. SAP*, Vol. 9, pp. 196–200, March 2001.
- [11] M. D. Skowronski and J. G. Harris, “Increased MFCC Filter Bandwidth for Noise-Robust Phoneme Recognition”, in *Proc. ICASSP-02*, Florida, May 2002.
- [12] H. Tolba and D. O’Shaughnessy, “Automatic Speech Recognition Based on Cepstral Coefficients and a Mel-based Discrete Energy Operator”, in *Proc. ICASSP-98*, Seattle, May 1998.
- [13] S. Young et al., “The HTK Book (for HTK Version 3.2)”, URL: <http://htk.eng.cam.ac.uk>.