THE TEAGER ENERGY BASED FEATURE PARAMETERS FOR ROBUST SPEECH RECOGNITION IN CAR NOISE

Firas Jabloun

Electrical Engineering Dept.
Bilkent University
06533 Ankara Turkey

A. Enis Çetin

Electrical Engineering Dept.
Bilkent University
06533 Ankara Turkey

ABSTRACT

In this paper, a new set of speech feature parameters based on multirate signal processing and the Teager Energy Operator is developed. The speech signal is first divided into nonuniform subbands in mel-scale using a multirate filter-bank, then the Teager energies of the subsignals are estimated. Finally, the feature vector is constructed by log-compression and inverse DCT computation. The new feature parameters have a robust speech recognition performance in car engine noise which is low pass in nature.

1. INTRODUCTION

It is shown in [1–6] that speech can be modeled as a linear combination of AM-FM signals in some cases. Each resonance, or formant, is represented by an AM-FM signal of the form

$$s(t) = a(t)\ \cos[\phi(t)] = a(t)\ \cos[\int_0^t \omega_i(\tau)d\tau + \phi(0)]. \tag{1}$$

where $a(t)$ is a time varying amplitude signal and $\omega_i(t)$ is the instantaneous frequency given by $\omega_i(t) = d\phi(t)/dt$. This model allows the amplitude and resonance frequency to vary instantaneously within one pitch period. In [3–6], it is also shown that the Teager Energy Operator (TEO) can track the modulation energy and identify the instantaneous amplitude and frequency. The TEO is defined by

$$\Psi_c[s(t)] = [\dot{s}(t)]^2 - s(t)\ddot{s}(t). \tag{2}$$

where $\dot{s} = \frac{ds}{dt}$. In the case of AM-FM signal of Equation (1),

$$\Psi_c[s(t)] \approx a^2(t)\omega_i^2(t). \tag{3}$$

assuming that the bandwidth of $a(t)$ is much smaller than that of $\omega_i(t)$ [6].

The idea that Ψ_c is an energy measure is motivated by the fact that an undamped oscillator consisting of a mass m and a spring of constant k has a displacement $x(t) = A\cos(\omega_0 t + \theta)$, with $\omega_0 = \sqrt{k/m}$. The instantaneous energy E_0 of this undamped oscillator is the sum of its kinetic and potential energies and equals the constant

$$E_0 = \frac{m}{2}(A\omega_0)^2. \tag{4}$$

In this case, $\Psi_c[x(t)] = (A\omega_0)^2$. So the energy of the linear oscillator is proportional to $\Psi_c[x(t)]$ [6].

In this paper, new feature parameters based on the nonlinear model of (1) are developed using the TEO. The speech signal is first divided into nonuniform subbands in mel-scale using a multirate filter bank. Then, in each subband, the Teager energies are estimated. Finally, the feature vector is constructed by log-compression and inverse DCT computation.

The idea behind using TEO instead of the commonly used instantaneous energy, is to take advantage of the modulation energy tracking capability of the TEO. This leads to a better representation of the formant information in the feature vector compared to the MELCEP [7] and SUBCEP [8] parameters in which the regular instantaneous energy is used.

In Section 2 we formally define the TEOCEP features and in Section 3 we present some properties of the TEO. In Section 4, we use the new parameters for speech recognition under car engine noise which is of low pass nature. Since the modulation energy of the car noise is very low compared to that of the speech signal, the TEOCEP's show better recognition performance than MELCEP's and SUBCEP's.

2. THE TEOCEP FEATURE PARAMETERS

In our method, multirate subband decomposition [8–10], is used in a tree structure to divide the speech signal $s(n)$ according to the mel-scale as shown in Fig.

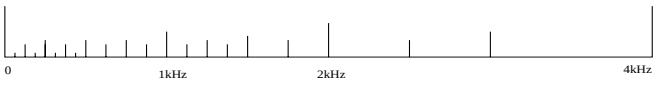

Figure 1: The sub-band frequency decomposition of the speech signal

(1), and 21 sub-signals $s_l(n)$, $l = 1, \ldots, L = 21$, are obtained. The filter bank of a biorthogonal wavelet transform is used in the analysis [11]. The lowpass filter has the transfer function

$$H_l(z) = \frac{1}{2} + \frac{9}{32}(z^{-1} + z^1) - \frac{1}{32}(z^{-3} + z^3). \quad (5)$$

and the corresponding high-pass filter has the transfer function

$$H_h(z) = \frac{1}{2} - \frac{9}{32}(z^{-1} + z^1) + \frac{1}{32}(z^{-3} + z^3). \quad (6)$$

For every sub-signal, the average Teager energy e_l

$$e_l = \frac{1}{N_l} \sum_{n=1}^{N_l} |\Psi_d[s_l(n)]| \; ; l = 1, \ldots, L. \quad (7)$$

is estimated. In (7), N_l is the number of samples in the l^{th} band, and $\Psi_d[.]$ is the discrete-time version of the continuous-time TEO which is obtained by approximating derivatives with the two-sample backward (or forward) difference $[s(n) - s(n-1)]/T$ where T is the sampling period. Without any loss of generality, T can be set to one, and the discrete-time version of the TEO is given by

$$\Psi_d[s(n)] = s^2(n) - s(n+1)s(n-1). \quad (8)$$

In this paper, the discrete version is used so from now on the subscript 'd' is dropped.

Although it is possible that the instantaneous Teager energy have negative values in very rare circumstances, the average value e_l is a positive quantity for most natural signals [4, 12]. Nonetheless, the magnitude of the Teager energy is used to ensure the non-negativity of e_l. Log compression and inverse DCT computation is finally applied to obtain the TEO-based cepstrum coefficients,

$$TC(k) = \sum_{l=1}^{L} \log(e_l) \cos[\frac{k(l-0.5)\pi}{L}] \; ; \; k = 1, \ldots, N. \quad (9)$$

We call the new features TEOCEP's. The first 12 $TC(k)$ coefficients are used in the feature vector. Twelve more coefficients obtained from the first-order differentials are also appended. A final feature vector with dimension 24 is obtained and is used for training and recognition.

The SUBCEP parameters used in [8] differ from the TEOCEP's just in the definition of the energy measure used in Equation (7). In [8],

$$\varepsilon_l = \frac{1}{N_l} \sum_{n=1}^{N_l} |s_l(n)| \; ; \; l = 1, \ldots, L \quad (10)$$

is used instead of e_l.

It is shown that the SUBCEP's perform slightly better than the well-known MELCEP features [8–10]. For this reason, the performance of the TEOCEP's are evaluated with respect to that of SUBCEP's.

3. PROPERTIES OF THE TEAGER ENERGY OPERATOR

The TEO is an efficient tool for nonlinear speech processing as the speech is composed of a superposition of AM-FM signals. To examine the behaviour of the TEO in the presence of noise, we calculate the mean of $\Psi[s(n)]$ or simply $\Psi_s(n)$

$$E\{\Psi_s(n)\} = E\{s^2(n)\} - E\{s(n+1)s(n-1)\} \quad (11)$$

Assuming that the speech is stationary within the current frame,

$$E\{\Psi_s(n)\} = R_s(0) - R_s(2). \quad (12)$$

where $R_s(k)$ is the autocorrelation function of $s(n)$.

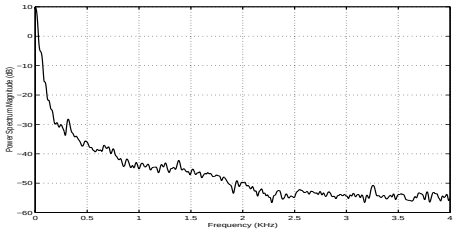

Figure 2: Power Spectrum Density of the car noise signal recorded inside a Volvo 340 on a rainy asphalt road by the *Institute for Perception-TNO, The Netherlands*

In this paper, we are interested in voice dialing applications and consider the colored car engine noise. The spectrum of the car noise $v(n)$ is mostly concentrated in low frequencies as shown in Figure 2. Thus, its correlation function varies very smoothly and it is almost flat near the origin for several lags. For this noise signal, the first three autocorrelation lags are estimated as

$$\begin{aligned} R_v(1) &= 0.9997 R_v(0) \\ R_v(2) &= 0.9991 R_v(0) \end{aligned} \quad (13)$$

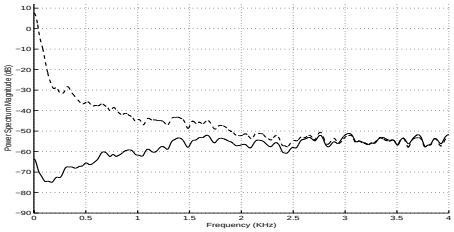

Figure 3: Spectrum of the car noise $v(n)$ (dashed line) and the spectrum of the Teager energy $\Psi[v(n)]$ (continuous line)

Since $R_v(0) \approx R_v(1) \approx R_v(2)$, we have $\Psi[v(n)] \approx 0$. This leads to the spectrum of $\Psi[v(n)]$ shown in Figure 3, which is almost flat and negligible compared to the spectrum of the noise $v(n)$.

Clearly, for a typical speech signal, s(n), the first three autocorrelation lags are not as close as in the car engine noise case. For example

$$\begin{aligned} R_s(1) &= 0.7415\ R_s(0) \\ R_s(2) &= 0.4584\ R_s(0) \end{aligned} \tag{14}$$

for the author's /a/.

Let the observed signal be $x(n) = s(n)+v(n)$, where $s(n)$ is the noise free speech signal and $v(n)$ is a zero mean additive noise.

The Teager energy of the noisy speech signal $x(n)$ is given by

$$\Psi[x(n)] = \Psi[s(n)] + \Psi[v(n)] + 2\widetilde{\Psi}[s(n), v(n)] \tag{15}$$

where $\widetilde{\Psi}_{[s(n),v(n)]=s(n)v(n)-\frac{1}{2}s(n-1)v(n+1)-\frac{1}{2}s(n+1)v(n-1)}$, is the cross-$\Psi$ energy of $s(n)$ and $v(n)$.

Since $s(n)$ and $v(n)$ are zero mean and independent, then the expected value of their cross-Ψ energy is zero. Moreover, $\Psi[v(n)]$ is negligible if the speech resonance frequency fall within the current analysis band [3]. Therefore

$$E\{\Psi[x(n)]\} \approx E\{\Psi[s(n)]\} \tag{16}$$

On the other hand, with the commonly used instantaneous energy, the noise bias persists and is proportional to the noise energy,

$$E\{x^2(n)]\} = R_s(0) + R_v(0) \tag{17}$$

As discussed in Section 2, TEOCEP's are obtained via multiresolution analysis. If a speech formant falls within an analysis band then its Teager energy is much higher than the Teager energy of the noise. Due to this reason, the formant information is well represented in the TEOCEP feature set.

4. SIMULATION RESULTS

A continuous density Hidden Markov Model based speech recognition system with 5 states and 3 Gaussian mixture densities is used in simulation studies. The recognition performances of the TEOCEP feature parameters are evaluated using the *TI-20* speech database of *TI-46 Speaker Dependent Isolated Word Corpus* which is corrupted by various types of additive noise. The *TI-20* vocabulary consists of ten English digits and ten control words. The data is collected from 8 male and 8 female speakers. There are 26 utterances of each word from each speaker, where 10 designated as training tokens and 16 designated as testing tokens.

SNR (dB)	TEOCEP	SUBCEP
30	99.66	99.15
10	99.26	99.05
7	99.37	97.98
5	99.05	97.02
3	98.84	96.41
0	98.17	95.14
-3	97.83	93.12
-5	96.86	90.62

Table 1: The average recognition rates of speaker dependent isolated word recognition system with SUBCEP and TEOCEP features for various SNR levels with Volvo noise recording.

Speaker dependent isolated word speech recognition simulations are described in Table 1 and Table 2 for Volvo car noise and white noise, respectively. The car noise is recorded inside a Volvo 340 on a rainy asphalt road by the *Institute for Perception-TNO, The Netherlands*. In the car noise case, the superiority of the TEOCEP's over the SUBCEP's is obvious especially at low SNR values. However, in white noise, just a slight improvement is achived at low SNR values. This can be theoretically predicted because for white noise $v(n)$, the autocorrelation function $R_v(k) = 0$ for $k \neq 0$.

In Table 3, speaker independent experiment results with the Volvo car noise are shown. The utterances of five men and five women were used for training. The utterances of the rest speakers are used to test the performance of the system. Again the TEOCEP parameters outperform the SUBCEP's especially at low SNR's.

5. CONCLUSION

In this paper, new feature parameters, TEOCEP's, for speech recognition are introduced. The new features

SNR (dB)	TEOCEP	SUBCEP
20	97.79	98.37
10	87.07	87.7
7	86.12	85.17
5	82.97	81.70
3	79.83	79.50

Table 2: The average recognition rates of speaker dependent isolated word recognition system with SUBCEP and TEOCEP features for various SNR levels with white noise.

SNR (dB)	TEOCEP	SUBCEP
30	91.22	91.25
10	91.13	90.96
7	90.74	89.94
3	89.10	88.40
0	87.13	86.63
-3	85.26	80.17

Table 3: The average recognition rates of speaker independent isolated word recognition system with SUBCEP and TEOCEP features for various SNR levels with Volvo noise recording.

are based on the Teager Energy Operator and the multirate sub-band analysis providing a robust recognition performance under car noise.

6. REFERENCES

[1] H. M. Teager, "Some observations on oral air flow during phonation," *IEEE Trans. on Speech and Audio Processing*, October. 1980.

[2] H. M. Teager and S. M. Teager, "Evidence for nonlinear speech production mechanisms in the vocal tract," *NATO Advanced Study Institute on Speech Production and Speech Modelling, Bonas, France*, July 1989.

[3] A. C. Bovik, P. Maragos, and T. Quatieri, "AM-FM energy detection and separation in noise using multiband energy operators," *IEEE Trans. on Signal Processing*, vol. 41, pp. 3245–3265, December 1993.

[4] P. Maragos, J. F. Kaiser, and T. Quatieri, "Energy separation in signal modulations with application to speech analysis," *IEEE Trans. on Signal Processing*, vol. 41, pp. 3025–3051, October 1993.

[5] P. Maragos, "Modulation and Fractal Models for Speech Analysis and Recognition," *Proceedings of COST-249 Meeting*, Feb. 1998.

[6] P. Maragos, T. Quatieri, and J. F. Kaiser, "On amplitude and frequency demodulation using energy operators," *IEEE Trans. on Signal Processing*, vol. 41, pp. 1532–1550, April 1993.

[7] S. B. Davis and P. Mermelstein, "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 28, pp. 357–366, August 1980.

[8] E. Erzin, A. Çetin, and Y. Yardımcı, "Subband analysis for robust speech recognition in the presence of car noise," *Proc. of the Int. Conf. on Acoustics, Speech and Signal Processing 1995 (ICASSP '95)*, May 1995.

[9] R. Sarikaya, B. L. Pellom, and J. H. Hansen, "Wavelet Packet Transform Features with Application to Speaker Identification," *NORSIG'98*, pp. 81–84, 1998.

[10] R. Sarikaya and J. N. Gowdy, "Subband Based Classification of Speech Under Stress," *Proc. of the Int. Conf. on Acoustics, Speech and Signal Processing 1998 (ICASSP '98)*, vol. 1, pp. 596–572, 1998.

[11] C. W. Kim, R. Ansari, and A. E. Çetin, "A class of linear-phase regular biorthogonal wavelets," *Proc. of the Int. Conf. on Acoustics, Speech and Signal Processing 1992 (ICASSP '92)*, vol. IV, pp. 673–677, 1992.

[12] A. C. Bovik and P. Maragos, "Conditions for positivity of an energy operator," *IEEE Trans. on Signal Processing*, Feb 1994.