Teager Energy Based Feature Parameters for Speech Recognition in Car Noise

Firas Jabloun, Student Member, IEEE, A. Enis Çetin, Senior Member, IEEE, and Engin Erzin, Member, IEEE

Abstract—In this letter, a new set of speech feature parameters based on multirate signal processing and the Teager energy operator is introduced. The speech signal is first divided into nonuniform subbands in mel-scale using a multirate filterbank, then the Teager energies of the subsignals are estimated. Finally, the feature vector is constructed by log-compression and inverse discrete cosine transform (DCT) computation. The new feature parameters have robust speech recognition performance in the presence of car engine noise.

Index Terms—Mel-scale, multirate signal processing, speech recognition, Teager energy operator.

I. INTRODUCTION

In this paper, a new set of speech feature parameters is proposed. The new parameters are developed using multirate signal processing and the Teager energy operator (TEO), which has been successfully used in various speech processing applications [1]–[5]. It is experimentally observed that the TEO can suppress the car engine noise, which makes the new feature parameters a good candidate for voice dialing systems in automobiles.

In continuous-time, the TEO is defined as

\[ \Psi_c[s(t)] = [s(t)]^2 - s(t)\dot{s}(t) \]  
(1)

where \( s(t) \) is a continuous-time signal and \( \dot{s} = ds/dt \). In discrete-time, the TEO can be approximated by

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

where \( s(n) \) is a discrete-time signal. In this work, the discrete-time version is used, and the subscript “d” is dropped from now on. Let \( s(n) \) be a discrete-time wide-sense stationary random signal. In this case

\[ E\{\Psi[s(n)]\} = E\{s^2(n)\} - E\{s(n+1)s(n-1)\} \]  
(3)

or

\[ E\{\Psi[s(n)]\} = R_s(0) - R_s(2) \]  
(4)

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

In general, the car engine noise, \( \nu(n) \), is mostly lowpass in nature. A typical example is shown in Fig. 1. For this noise signal, the relation between the first three autocorrelation lags are estimated as \( R_v(1) = 0.9997R_v(0) \) and \( R_v(2) = 0.9991R_v(0) \). Since \( R_v(0) \approx R_v(1) \approx R_v(2) \) we have \( E[\Psi[\nu(n)]] \approx 0 \). Due to this reason, the spectrum of \( \Psi[\nu(n)] \) shown in Fig. 2 is almost negligible compared to the spectrum of the noise \( \nu(n) \).

For a typical speech signal, \( s(n) \), the first three autocorrelation lags are not as close to each other. For example, \( R_s(1) = 0.7415R_s(0) \) and \( R_s(2) = 0.4584R_s(0) \) for the first author’s /a/. \( R_v(1) = 0.97R_v(0) \), \( R_v(2) = 0.91R_v(0) \) for the second author’s /f/. \( R_s(1) = 0.84R_s(0) \), and \( R_s(2) = 0.73R_s(0) \), for the second author’s /s/.

In practice, the observed signal is the sum of the speech signal and the noise. Let the observed signal be \( x(n) = s(n) + \nu(n) \) where \( s(n) \) is the noise-free speech signal and \( \nu(n) \) is a zero mean additive noise, which is independent from \( s(n) \). The Teager energy of the noisy speech signal \( x(n) \) is given by

\[ \Psi[x(n)] = \Psi[s(n)] + \Psi[\nu(n)] + 2\Psi[s(n)\nu(n)] \]  
(5)

where \( \Psi[s(n)] \) and \( \Psi[\nu(n)] \) are the cross-\( \Psi \) energy of \( s(n) \) and \( \nu(n) \). Since \( s(n) \) and \( \nu(n) \) are zero mean and independent the expected value of their cross-\( \Psi \) energy is zero.

Thus, \( E[\Psi[x(n)]] = E[\Psi[s(n)]] + E[\Psi[\nu(n)]] \). Furthermore, \( E[\Psi[\nu(n)]] \) is negligible compared to \( E[\Psi[s(n)]] \) for

Manuscript received December 3, 1998. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. J. H. L. Hansen.

F. Jabloun and A. E. Çetin are with the Department of Electrical and Electronics Engineering, Bilkent University, Ankara 06533, Turkey (e-mail: cetin@ee.bilkent.edu.tr).

E. Erzin is with Lucent Technologies, Whippany, NJ 07981 USA (e-mail: erzin@lucent.com).

Publisher Item Identifier S 1070-9908(99)07909-2.
the car engine noise, i.e.,

\[
E\{\Psi[x(n)]\} \approx E\{\Psi[s(n)]\},
\]  

(6)

Hence, the effect of car engine noise can be eliminated by using the TEO in feature extraction. On the other hand, the commonly used energy has no filtering capability because

\[
E\{x^2(n)\} = R_s(0) + R_w(0).
\]  

(7)

For this reason, we expect a TEO-based feature set to produce better recognition rates than the regular energy based features in car engine noise.

In Section II, new Teager energy operator based cepstral (TEOCEP) feature parameters are formally defined. In order to obtain the TEOCEP parameters, the speech signal is first divided into nonuniform subbands in mel-scale using a multirate filterbank. Then the Teager energies are estimated in each subband and the feature vector is constructed by log-compression and inverse discrete cosine transform (IDCT) computation. In Section III, the new parameters are used in isolated word recognition under car engine noise and it is experimentally observed that the TEOCEP parameters produce better recognition performance than MELCEP’s [6] and subband decomposition based cepstral (SUBCEP) parameters.

### II. THE TEOCEP FEATURE PARAMETERS

In our method, multirate subband decomposition [7]–[9] is used in a tree structure to divide the speech signal \(x(n)\) according to the mel-scale as shown in Fig. 3, and 21 subsignals \(x_l(n), l = 1, \cdots, L = 21\), are obtained. The filter bank corresponding to a biorthogonal wavelet transform is used in the analysis [10]. The lowpass \(H_{0,l}(z)\) and highpass \(H_{1,l}(z)\) filters have the transfer functions

\[
H_{0,l}(z) = \frac{1}{2} \pm \frac{\sqrt{2}}{32} (z^{-1} + z^1) \pm \frac{1}{32} (z^{-3} + z^3)
\]  

(8)

respectively. For every subsignal, the average Teager energy \(c_l\)

\[
c_l = \frac{1}{N_l} \sum_{n=1}^{N_l} |\Psi[x_l(n)]|; \quad l = 1, \cdots, L
\]  

(9)

is estimated. In (9), \(N_l\) is the number of samples in the \(l\)th band. Due to downsampling operations in multirate subband decomposition \(N_l\) values are less than the number of samples, \(N\), in a speech frame: \(N_l = N/16, l = 1, 2, \cdots, 18; N_l = N/8, l = 19, 20\), and \(N_l = N/4, l = 21\). In our simulation studies, the frame size is chosen as 48 ms, which is equivalent to \(N = 384\) samples at 8 kHz sampling rate, and the overlap between the frames is 32 ms.

Although it is possible that the instantaneous Teager energy has a negative value in very rare circumstances, the average value \(c_l\) is a positive quantity for most natural signals [4], [11] as \(R_s(0)\) is usually larger than \(R_w(0)\). Nonetheless, the magnitude of the Teager energy is used in (9) to ensure the nonnegativity of \(c_l\).

At the last step, log compression and IDCT computation is applied to obtain the TEO-based cepstrum coefficients

\[
TC(k) = \sum_{l=1}^{L} \log(c_l) \cos\left[\frac{k(l - 0.5)\pi}{L}\right] \quad k = 1, \cdots, N,
\]  

(10)

We call the new feature set TEOCEP parameters. 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 [7] differ from the TEOCEP parameters in the definition of the energy measure used in (9). In [7], \(\ell_1\) energy

\[
\varepsilon_l = \frac{1}{N_l} \sum_{n=1}^{N_l} |x_l(n)|; \quad l = 1, \cdots, L
\]  

(11)

is used instead of \(c_l\).

It is shown that the SUBCEP parameters perform slightly better than the well-known MELCEP parameters [7]–[9]. For this reason, the performance of the TEOCEP parameters are evaluated with respect to that of SUBCEP parameters.

### III. SIMULATION RESULTS

A continuous density hidden Markov model based speech recognition system with five states and three 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 eight male and eight female speakers. There are 26 utterances of each word from each speaker of which ten are designated as training tokens and 16 designated as testing tokens.

Speaker-dependent isolated word speech recognition simulations are presented in Tables I–III. In the first two tables car noise is added on the speech signal and in Table III the speech signal is corrupted by additive white noise. The first car noise is recorded inside a Volvo 340 on a rainy asphalt road by the Institute for Perception-TNO, The Netherlands. The spectrum of this noise signal is shown in Fig. 1. The second set of results in Table II is obtained for the noise recorded inside a Mazda 626 on an asphalt road traveling at 90 km/h (55 miles/h). The same filterbank is used to generate the SUBCEP and TEOCEP parameters. The frame size is chosen as 48 ms with an overlap of 32 ms. In the car noise case, the superiority of the TEOCEP parameters over the SUBCEP parameters is obvious, especially at low signal-to-noise ratio (SNR) values. However, in white noise, just a slight improvement is achieved at low SNR values. This is expected because for white noise, the TEO does not perform any filtering.

In Table IV, 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 of the speakers are used to test the performance of the system. Again, the TEOCEP parameters outperform the SUBCEP parameters especially at low SNR’s.

REFERENCES


### TABLE I

<table>
<thead>
<tr>
<th>SNR (dB)</th>
<th>TEOCEP</th>
<th>SUBCEP</th>
</tr>
</thead>
<tbody>
<tr>
<td>30</td>
<td>99.66</td>
<td>99.15</td>
</tr>
<tr>
<td>10</td>
<td>99.26</td>
<td>99.05</td>
</tr>
<tr>
<td>7</td>
<td>99.37</td>
<td>97.98</td>
</tr>
<tr>
<td>5</td>
<td>99.05</td>
<td>97.02</td>
</tr>
<tr>
<td>3</td>
<td>98.84</td>
<td>96.41</td>
</tr>
<tr>
<td>0</td>
<td>98.17</td>
<td>95.14</td>
</tr>
<tr>
<td>-3</td>
<td>97.83</td>
<td>93.12</td>
</tr>
<tr>
<td>-5</td>
<td>96.86</td>
<td>90.62</td>
</tr>
</tbody>
</table>

### TABLE II

<table>
<thead>
<tr>
<th>SNR (dB)</th>
<th>TEOCEP</th>
<th>SUBCEP</th>
</tr>
</thead>
<tbody>
<tr>
<td>30</td>
<td>99.54</td>
<td>99.43</td>
</tr>
<tr>
<td>10</td>
<td>99.41</td>
<td>99.10</td>
</tr>
<tr>
<td>7</td>
<td>99.12</td>
<td>98.30</td>
</tr>
<tr>
<td>5</td>
<td>99.11</td>
<td>97.17</td>
</tr>
<tr>
<td>3</td>
<td>98.93</td>
<td>96.83</td>
</tr>
<tr>
<td>0</td>
<td>98.05</td>
<td>95.20</td>
</tr>
<tr>
<td>-3</td>
<td>97.81</td>
<td>93.05</td>
</tr>
<tr>
<td>-5</td>
<td>96.57</td>
<td>90.54</td>
</tr>
</tbody>
</table>

### TABLE III

<table>
<thead>
<tr>
<th>SNR (dB)</th>
<th>TEOCEP</th>
<th>SUBCEP</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>97.79</td>
<td>98.37</td>
</tr>
<tr>
<td>10</td>
<td>87.07</td>
<td>87.70</td>
</tr>
<tr>
<td>7</td>
<td>86.12</td>
<td>85.17</td>
</tr>
<tr>
<td>5</td>
<td>82.97</td>
<td>81.70</td>
</tr>
<tr>
<td>3</td>
<td>79.83</td>
<td>79.50</td>
</tr>
</tbody>
</table>

### TABLE IV

<table>
<thead>
<tr>
<th>SNR (dB)</th>
<th>TEOCEP</th>
<th>SUBCEP</th>
</tr>
</thead>
<tbody>
<tr>
<td>30</td>
<td>91.22</td>
<td>91.25</td>
</tr>
<tr>
<td>10</td>
<td>91.13</td>
<td>90.96</td>
</tr>
<tr>
<td>7</td>
<td>90.74</td>
<td>89.94</td>
</tr>
<tr>
<td>3</td>
<td>89.10</td>
<td>88.40</td>
</tr>
<tr>
<td>0</td>
<td>87.13</td>
<td>86.63</td>
</tr>
<tr>
<td>-3</td>
<td>85.26</td>
<td>80.17</td>
</tr>
</tbody>
</table>