NOISE SUPPRESSION BASED ON TEAGER ENERGY OPERATOR FOR IMPROVING THE ROBUSTNESS OF ASR FRONT-END

Zhao Junhui, Kuang Jingming, Xie Xiang, Huang Shilei

Research Center of Digits Communication Technologies, Beijing Institute of Tech, China
{zkes1000, xiexiang, huang_shilei}@bit.edu.cn

ABSTRACT

In this paper, we proposed a new noise suppression method based on Teager Energy Operator in advancing the noise robustness of speech recognition front-end. The presented method attempts to remove a distortion estimation in Teager energy domain, especially, a Teager energy estimation of noise signal is subtracted from the noisy speech signal. This approach differs significantly from the traditional spectral subtraction, which is frequency domain based, and we use it in this work as a complementary technique to the Teager energy based feature parameters [1]. A mandarin digit string recognition task is performed for evaluating the performance of the proposed method. The recognition results show a robust speech recognition performance in noisy environment.

1 INTRODUCTION

There is much experimental and theoretical evidence for the existence of amplitude and frequency modulation (AM-FM) in speech signals, which make the amplitude and frequency of the resonance (formant) vary instantaneously within a pitch period [2-7]. Motivated by this evidence, Maragos and Kaiser [8] model each speech resonance with an AM-FM signal of the form

$$s(t) = a(t)\cos[2\pi\int_0^t f(\tau)d\tau] \tag{1}$$

and the total speech signal as a superposition of such AM-FM signals. Here *a(t)* and *f(τ)* are the instantaneous amplitude and frequency which represent the time-varying formant signal. It is shown in [4-8] that the Teager Energy Operators (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 *c* corresponds to continuous signal and $\dot{s} = ds/dt$.

Jabloun and Cetin [1] have proposed a new feature parameters based on the TEO, which is defined as TEOCEP. The speech signal is first divided into nonuniform subbands in mel-scale. Then, in each band, the Teager energies are estimated. Finally, the feature vector is constructed by log-compression and inverse DCT computation. The new feature vector can provide a robust recognition performance under car noise that is mostly concentrated in low frequencies.

However, the TEOCEP can not bring satisfied recognition accuracy in other noisy environments, especially non-stationary noise. In this paper, we present a new noise suppression algorithm based on TEOCEP for improving the robustness of speech recognition system. It attempts to obtain an estimation of the Teager energy of clean speech, by removing a distortion caused by noise signal from the Teager energy of the observed signal. We develop the method as a complementary approach to TEOCEP feature.

The paper is organized as follows. Section 2 gives a description of TEOCEP feature and section 3 explains the presented noise suppression algorithm. Section 4 shows recognition experiments. In section 5 some conclusions are given.

2 THE TEOCEP FEATURE PARAMETERS

The scheme of TEOCEP feature is given in Fig 1. The extra box in dashed line will be explained in the next section. In our case, 26 triangular bandpass filters ranging from 200Hz to 4k Hz are used to divide the input speech signal according to the mel-scale. For each sub-signal, the average Teager energy e_l is calculated.

$$e_l = \frac{1}{N}\sum_{n=1}^{N}\left|\Psi_d[s_l(n)]\right|; l = 1,...L \tag{3}$$

here *N* is the frame length and *l* is filter bank index and $\Psi_d[s(n)]$ is the discrete time version of the TEO which is defined by

$$\Psi_d[s(n)] = s^2(n) - s(n+1)s(n-1) \tag{4}$$

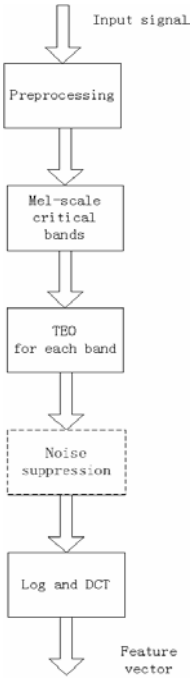

Figure 1. Block diagram of the TEOCEP feature

In this paper, the discrete version is used so from now on the subscript 'd' is dropped. The cepstral coefficients are found with inverse DCT computation on the logarithm of the output Teager energies of the filterbank. A vector size of 39 is used for the training and recognition which is obtained from 13 cepstral coefficients and the corresponding delta and acceleration coefficients.

3 NOISE SUPPRESSION BASED ON TEO

3.1 Basic Idea

Let us consider a speech signal *s(n)* degraded by uncorrelated additive noise *v(n)*. The resulting signal is then

$$y(n) = s(n) + v(n) \tag{5}$$

The Teager energy of the noisy speech signal $\Psi[y(n)]$ is given by

$$\Psi[y(n)] = \Psi[s(n)] + \Psi[v(n)] + 2\tilde{\Psi}[s(n), v(n)] \tag{6}$$

Where $\Psi[s(n)]$ and $\Psi[v(n)]$ are respectively the Teager energy of the speech signal and the additive noise. $\tilde{\Psi}[s(n), v(n)]$ is the cross-Ψ energy of *s(n)* and *v(n)*, such that

$$\tilde{\Psi}[s(n), v(n)] = s(n)v(n) - 0.5s(n-1)v(n+1) - 0.5s(n+1)v(n-1) \tag{7}$$

We need to obtain an estimation of $\Psi[s(n)]$ which is the Teager energy of the clean speech. In Equation (6), only $\Psi[y(n)]$ can be directly derived from the observed data. However $\Psi[v(n)]$ and $\tilde{\Psi}[s(n), v(n)]$ can be approximated respectively by $E[\Psi[v(n)]]$ and $E[\tilde{\Psi}[s(n), v(n)]]$, where E[.] denote the ensemble average. With the hypothesis that the noise *v(n)* is uncorrelated with speech *s(n)*, we have the expected value of their cross-Ψ energy $\tilde{\Psi}[s(n), v(n)]$ equal to zero and we can thus derive an estimation $\hat{\Psi}[s(n)]$ of $\Psi[s(n)]$

$$\hat{\Psi}[s(n)] = \Psi[y(n)] - E[\Psi[v(n)]] \tag{8}$$

We call the new algorithm as TES (Teager Energy Subtraction). The form of our presented method is similar with the spectrum subtraction, which is frequency domain based. However, it is very different that the new algorithm works in Teager energy domain which is derived from TEOCEP feature vector.

3.2 Implementation

The Teager energy subtraction needs an estimation of the noise Teager energy. This estimation can be computed during non-speech period; however it can not track the changes of noise signal. In our case, the noise Teager energy is updated with a first order recursion process and an adaptive threshold is used to stop the recursion when the speech is most likely to be present. For the *l*th band in frame *i*, an estimation $\Psi[v_l(i)]$ of the noise Teager energy is obtained by the first order recursion (from here we ignore the time subscript *n* and start to use the frame index *i*)

$$\begin{aligned} &if\ \Psi[y_l(i)] \le \beta\Psi[v_l(i-1)] \\ &\quad then\ \Psi[v_l(i)] = \alpha\Psi[v_l(i-1)] + (1-\alpha)\Psi[y_l(i)] \\ &else\ \Psi[v_l(i)] = \Psi[v_l(i-1)] \end{aligned} \tag{9}$$

Where β is very similar to traditional signal to noise ratio (SNR) which has an optimal value of 2. A typical value for α is 0.99, which corresponding to an adaptation over 100 frames, that is 1 second. The initialization of the noise Teager energy estimation is done on the first 10 frames (this makes the assumption that the first 10 frames contain only noise).

It is important to note that Equation (8) does not guarantee that $\hat{\Psi}[s]$ is positive. Negative value should be set to zero or better to a constant non-zero minimum value. Since the TEO leads to a better representation of the formant information, the Teager energy of the speech is much higher than the noise's, especially when a speech formant falls within an analysis band. In this work, we apply the following method to prevent the negative value. If we define $T[s_l(i)]$ as

$$T[s_l(i)] = \Psi[y_l(i)] - \Psi[v_l(i)] \quad (10)$$

Then for each band, the Teager energy of clean speech is given by

$$\hat{\Psi}[s_l(i)] = \begin{cases} T[s_l(i)] & if T[s_l(i)] > \lambda\Psi[v_l(i)] \\ \lambda\Psi[v_l(i)] & otherwise \end{cases} \quad (11)$$

λ defines the minimum Teager energy value after subtraction.

4 RECOGNITION EXPERIMENTS

4.1 Experiment Setup and Database

A speaker-independent Mandarin digits string recognition task is performed for evaluating the performance of the proposed algorithm. 11 Mandarin digits ranging from zero to nine including "yao1" (equivalent to "one") are recorded by 63 male and 62 female speakers at 8k Hz, 16 bits. The training data consists of 4766 utterances and recognition is performed on other 2170 utterances. The recognition engine throughout our experiments is based on the HTK software package version 3.1 [9]. The digits are modeled as whole word HMMs with the following parameters:

- 8 states per word (according to 10 states in HTK notation with 2 dummy states at beginning and end)
- simple left-to-right models without skips over states
- mixture of 3 Gaussians per state
- only the variances of all acoustic coefficients (No full covariance matrix).

The NOISEX-92 noise database [10] was used to evaluate the noise robustness of the presented noise suppression algorithm. Babble, Pink and Factory noises are selected for this work. In training phase, the references are obtained on clean data only. The noise signal is added to the test utterances at SNRs of 20dB, 15dB, 10dB, 5dB and 0dB. Furthermore, the clean test utterances without adding noise is also taken into account.

4.2 Comparison between TEOCEP and TEOCEP+TES

The evaluation results of the TES algorithm performance under the three noise conditions are shown in Figure 2, 3 and 4, respectively. Table 1 summarized the results of these figures. The recognition results are given in percent of word error rate (WER).

The first point to note is that the TES can improve the robustness of speech recognition system under noisy environments. In Babble noise, the average WER of TEOCEP+TES is 21.43%, yielding a 26.9% improvement compared to 29.31% in TEOCEP. In Pink and Factory noise case the performance improvements can be achieved when the TES is applied.

Secondly, using TES can not help improve the recognition accuracy on clean or slightly noisy speech. In particularly, in clean case (without adding noise), the WER is increased from 5.67% to 6.54%. The recognition accuracy has not been changed in 20 dB of SNR for the three noises. Although the difference is not significant, it is can be concluded that the TES has no effect on speech with high SNR.

Finally, when the TES is performed on TEOCEP features, the performance improvement is obvious especially at low SNR values. With Pink noise at 0 dB of SNR the error rate is reduced from 73.9% to 64.56%. The effect is even more important when the noise is non-stationary, such as: in Babble noise, at 0 dB of SNR, the WER is decreased from 67.14% to 45.55% and in Factory noise, the WER is reduced from 74.9% to 54.56%.

Table 1. The average WER (%) with TEOCEP and TEOCEP+TES

	Babble noise	Pink Noise	Factory noise
TEOCEP	29.31	33.40	31.37
TEOCEP +TES	21.43	30.41	23.11

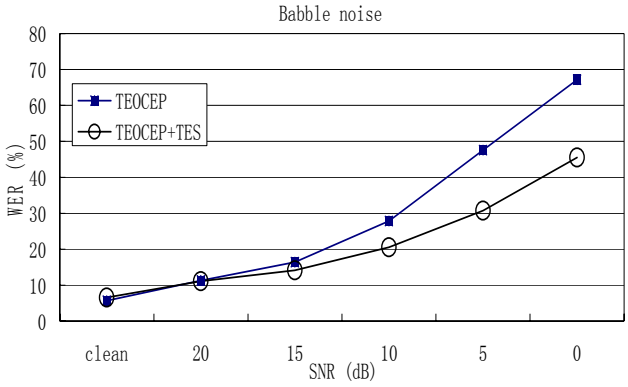

Figure 2. The recognition rates of Mandarin digits string recognition task with TEOCEP and TEOCEP+TES for various SNR levels in Babble noise

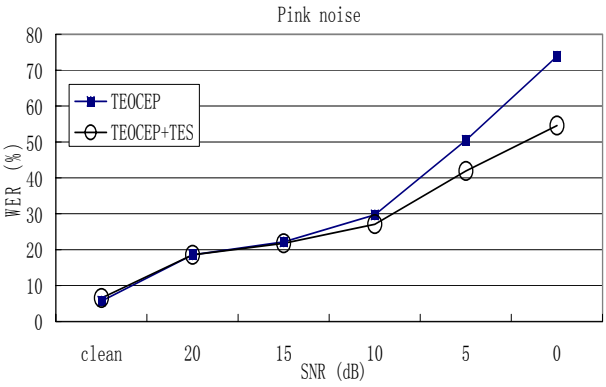

Figure 3. The recognition rates of Mandarin digits string recognition task with TEOCEP and TEOCEP+TES for various SNR levels in Pink noise

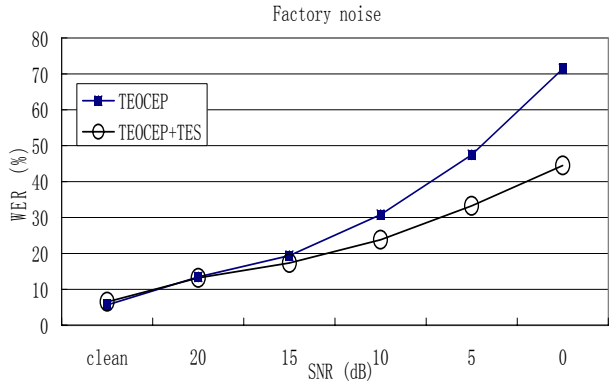

Figure 4. The recognition rates of Mandarin digits string recognition task with TEOCEP and TEOCEP+TES for various SNR levels in Factory noise

4.3 Comparison between TES and Spectral Subtraction

In this experiment, we evaluate the robust recognition performance of the proposed TES algorithm compared to the conventional Spectral Subtraction (SS). For SS case, the noisy speech signal is firstly enhanced by SS method and then the TEOCEP feature vector is extracted from the enhanced speech signal for comparison with the TEOCEP+TES. Considering that Factory noise is the least stationary and the most difficult for recognition, we present only the results obtained by Factory noise and similar conclusions as presented below can be derived for the other different noise sources. The recognition result in Factory noise is shown in Table 2.

Table 2. The WER (%) with TEOCEP +SS and TEOCEP+TES in Factory noise

SNR (dB)	20	15	10	5	0
TEOCEP+ SS	12.70	17.29	26.93	43.56	52.01
TEOCEP +TES	13.22	17.34	23.8	33.28	44.51

From Table 2, it turns out that for high SNR case, such as 20 dB and 15 dB, the performance of TES is no better than that of SS but their recognition accuracies are very close. With SNR dropped, TES can achieve better performance than SS. In particular, TES can outperform SS significantly at the worst SNR of 0 dB, which shows that TES is a more effective noise suppression method for improving the robustness of speech recognition system.

5 CONCLUSIONS

In this paper, we described a noise suppression algorithm based on Teager Energy Operator that is evaluated on a Mandarin digits string recognition task. The proposed method is derived from Teager energy subtraction. It can be considered as a complementary technology to the TEOCEP feature. Under three noise conditions, the results show that the TES algorithm can achieve a robust speech recognition performance.

6 ACKNOWLEDGEMENTS

The authors would like to thank Stephen dobler and Dongsheng Luo for several beneficial discussions. The research was supported by the corporation project of Sweden Ericsson Company and Beijing Institute of Technology.

7 REFERENCES

[1] Firas. Jabloun, A. Enis. Cetin, "The Teager Energy Based Feature Parameters for Robust Speech Recognition in Car Noise", *ICASSP1999*, vol.1, pp 273-276, 1999.

[2] H. M. Teager, "Some observations on oral air flow during phonation," *IEEE Trans. on Speech and Audio Proc*., Oct. 1980.

[3] 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.

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

[5] A. C. Bovik and P. Maragos. "Conditions for Positivity of an Energy Opterator", *IEEE Trans. On Signal Processing*, Fre 1994.

[6] A. C. Bovik, P. Maragos, and T. F. Quatieri, "AM-FM Energy Detection and Separation in Noise Using Multiband Energy Operators", *IEEE Trans. SignalProcessing*, vol. 41, Dec. 1993.

[7] D. Dimitriadis, P. Maragos, V. Pitsikalis and A. Potamianos, "Modulation and Chaotic Acoustic Features for Speech Recognition", *J. Control and Intelligent Systems*, Invited Paper, accepted for publication, 2002.

[8] P. Maragos, J. F. Kaiser, and T. F. Quatieri, "Energy Separation in Signal Modulations with Application to Speech Analysis", *IEEE Trans. Signal Processing*, vol. 41, pp. 3024–3051, Oct. 1993.

[9] S. Young, The HTK Book, Cambridge Research Lab: Entropics, Cambridge, England, 2001.

[10] A.Varga. "The NOISEX-92 study on the effect of additive noise on automatic speech recognition", Documentation included in the NOISEX-92 CD-ROMS, 1992.