TIME-FREQUENCY AVERAGING OF NOISE COMPENSATION FILTERS FOR ASR

Sunil Sivadas

Multimedia Technologies Laboratory, Nokia Research Center, Tampere, Finland.
sunil.sivadas@nokia.com

ABSTRACT

In this paper we show that time-frequency averaging of noise compensation filters, estimated using well-known techniques such as spectral subtraction, Wiener filtering and Ephraim-Malah approach, improves the performance of automatic speech recognition system in non-stationary noisy environments. The experiments were conducted on a multilingual isolated word recognition task in signal to noise ratio ranging from 5dB to 20dB and clean conditions.

1. INTRODUCTION

Susceptibility of automatic speech recognizers to environmental noise is a hindrance to their widespread usage in voice interfaces. In the case of mobile devices, depending on the ambience of the user, the noise could be background sounds like traffic noise, background speech or music. In all cases we deal with a difficult situation where only the noisy signal is available. No additional signal, for e.g. a second microphone to measure the ambient noise, is available to improve the quality of the corrupted signal.

Although there are various schemes for improving the Signal to Noise Ratio (SNR) of noisy speech, only a few offer both acceptable performance for real world noises and low complexity to be implemented on a mobile device. Among them are subtractive methods based on spectral subtraction and Wiener filtering [1][2]. However, all single channel subtractive-type algorithms are characterized by a tradeoff between the amount of noise reduction, the speech distortion, and the level of musical residual noise, which can be modified by varying the subtraction parameters. Algorithms are usually limited to the use of fixed optimized parameters, which are difficult to choose for all speech and noise conditions. Various methods have been proposed to reduce this effect: magnitude averaging [1], over-subtraction of noise and introduction of noise floor [3], soft-decision noise suppression filtering [4], optimal MMSE estimation of short-time spectral amplitude [5], nonlinear spectral subtraction [6] and applying properties of human auditory system [7].

Main cause of residual noise and distortion of estimated clean speech is the error in estimating various parameters in the noise suppression filter. The algorithm presented here addresses these issues through a simple time-frequency averaging of the subtractive filter. This paper is organized as follows. In Section 2, the principles of subtractive noise suppression algorithms and the proposed time-frequency averaging are described. In Section 3, experimental results are presented.

2. ALGORITHM DESCRIPTION

Consider a single channel system corrupted by additive background stationary noise. The noisy speech can be expressed as:

$$x(n) = s(n) + n(n)$$

where $s(n)$ is the original speech signal and $n(n)$ is the additive noise. Speech and noise are assumed to be uncorrelated. The processing is done on a frame-by-frame basis in the frequency domain. It is mainly composed of two phases: calculations of the smoothed subtractive filter and noise subtraction. Figure 1 shows a block diagram of the noise suppression scheme. Each of the blocks in the figure are explained in subsequent subsections.

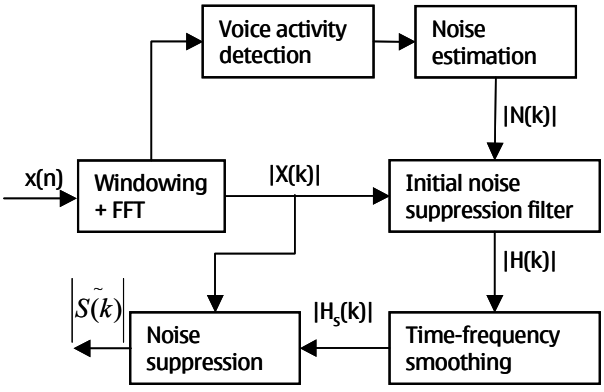

Figure 1. Block diagram of the noise suppression scheme

2.1. Spectral Amplitude Estimation

The clean speech magnitude spectrum is estimated using a time-varying linear filter dependent on the noisy signal spectrum and on the estimated noise spectrum [3].

$$|\hat{S}(k,n)| = |H(k,n)| * |X(k,n)| \qquad (1)$$

The filter gain, $|H(k,n)|$, can be estimated using many well-known techniques, such as spectral subtraction,

Wiener filter and Ephraim-Malah's MMSE approach. We use Wiener filter approach. The gain function is computed according to the following equation:

$$|H(k,n)| = \max\left(\frac{|X(k,n)|^2 - |\hat{N}(k,n)|^2}{|X(k,n)|^2}, \beta * |\hat{N}(k,n)|^2\right) \quad (2)$$

where, $|X(k,n)|$ is the k^{th} spectral component of frame number n computed using Short Term Fourier Transform (STFT) of the noisy speech and $|\hat{N}(k,n)|$ is the smoothed spectral component of noise signal. $|\hat{N}(k,n)|$ is given by

$$|\hat{N}(k,n)|^2 = \alpha|\hat{N}(k,n)|^2 + (1-\alpha)|X(k,n)|^2 \quad (3)$$

$|\hat{N}(k,n)|$ is updated only during speech pauses detected by a Voice Activity Detector (VAD). The scaling factor, β, of noise floor is set to 0.001 and the forgetting factor, α, for noise spectrum update is set to 0.99. Noise floor is introduced to limit the effect of residual noise at the expense of increased background noise. During the first 15 frames $|\hat{N}(k,n)|$ is computed as the running average of noisy signal spectrum $|X(k,n)|$.

$$\alpha = \frac{1}{n+1}, if\ n \le 15 \quad (4)$$

Figure 2(a) shows a speech waveform in background speech noise at 5dB SNR. In order to visualize the effect of the filtering scheme, a typical power spectrum of a noisy signal and filter transfer function are shown in Figure 2 (b) and Figure 2(c). Figure 2 (e) shows the estimated clean power spectrum in Equation (1). It can be clearly seen that the estimated clean power spectrum $|\hat{S}(k,n)|$ has little residual noise but some of the speech energy is also suppressed. This is due to the errors in the estimation of filter gain. In non-stationary noisy environments it is hard to keep track of the noise spectrum changes. For speech enhancement purposes minimizing residual noise at the expense of attenuation of speech portions maybe desirable. For ASR it is crucial to minimize the distortion in speech regions. There are various methods to minimize this. In the next section we present a simple solution.

2.2. Time-frequency averaging

The magnitude averaging solution presented in [1] has an inherent problem that the speech is nonstationary, and therefore the span of the averaging filter is limited. A weighted average of several frames is adopted in [8]. More recently, 2-D transforms have been used to capture the correlation [9]. It is clear that there is gain exploiting the correlation between adjacent frames. In this work we average the estimated filter in both time and frequency. A similar approach is presented in [10]. First the filter in Equation (2) is time averaged to obtain a new filter as given in the following Equation.

$$|H_{ts}(k,n)|^2 = \alpha_{filter} * |H_{ts}(k,n-1)|^2 + (1-\alpha_{filter}) * |H(k,n)|^2 \quad (5)$$

Next, the time-averaged filter is smoothed in frequency with a rectangular window.

$$|H_{tfs}(k,n)|^2 = \frac{1}{2*L+1}\sum_{l=-L}^{l=L}|H_{ts}(k+l,n)|^2 \quad (6)$$

The time and frequency averaged filter is applied to the power spectrum of the noisy signal in Equation (1) to obtain an estimate of the clean power spectrum. Figure 2(d) and Figure 2(f) show the filter gain and the estimated clean power spectrum. It can be observed that the resulting clean power spectral estimate has a lot more background noise in it than the estimate without averaging (Figure 2(e)), but the speech regions are well preserved in the new approach.

2.3. Cepstral normalisation

Mel-frequency cepstral coefficients (MFCC) are computed from the estimated clean magnitude spectrum. First and second derivatives are calculated from the static coefficients. They are then mean and variance normalized according to the following equation.

$$c_i'(n) = \frac{c_i(n) - m_i(n)}{\sigma_i(n) + \lambda} \quad (7)$$

where, $c_i(n)$ is the i^{th} cepstral feature at frame n, $m_i(n)$ and $\sigma_i(n)^2$ are the mean and the variance of the i^{th} cepstral feature estimated at frame n, and $c_i'(n)$ is the normalized cepstral feature at frame n. The value of bias λ is fixed at 1.0. All the cepstral features are normalized. The bias is introduced to smooth small estimates of variance.

3. EXPERIMENTAL RESULTS

The algorithms were tested on a multilingual small vocabulary isolated word recognition task. Test set comprises of approximately 40,000 words from seven European languages: Finnish, Swedish, German, English, Danish, Icelandic and Norwegian. The size of vocabulary per language was approximately 120.

The baseline front-end used in the experiments was based on 13 FFT-derived Mel-frequency cepstral coefficients (MFCC) and their first and second order derivatives (39 coefficients in total). Recursive mean removal was applied on all components of the resulting feature vectors, and the variance of only c_0 and its derivatives are normalized to unity [11].

The baseline acoustic model sets consists of 3 state monophone models with 8 Gaussian densities per state. The model sets were trained on an in-house training set containing clean speech data from various European

languages. Both sets contained a total of 75 multilingual phone models that were used to model the basic acoustic units of the seven European languages mentioned above.

The Word Error Rates (WER) of Wiener filter based front-end with time-frequency averaging are tabulated in Table 1. Noise is artificially added to the clean utterances at Signal to Noise Ratios (SNR) ranging from 5dB to 20dB in steps of 5dB. The noise waveform is created by concatenating car noise, background speech and music. The noise segment to be added to the clean speech utterance and the SNR is randomly selected. This makes it difficult to tune the performance of the front-end to a particular noise condition.

	WER (%)	
	Clean	**Noisy**
Baseline	3.49	12.53
Wiener filter (no averaging)	3.46	11.17
Wiener filter (time averaging)	3.51	10.89
Wiener filter (frequency averaging)	3.49	10.76
Wiener filter (time-frequency averaging)	3.58	10.22

Table 1. Performances of Wiener filter with time and frequency averaging.

It can be observed that the frequency averaging provides more gain than time averaging alone. The combined approach is the best combination. Further, we examine an alternative to rectangular window for frequency smoothing. Table 2 lists the performance, in noise only, using a triangular window for frequency averaging. We also examined the application of Wiener filter in the Mel-frequency domain. But none of these approaches perform as well as the rectangular window based averaging. We compared our approach with two other smoothing based noise robust front-ends, namely, ETSI Advanced front-end standard [12] and J-RASTA processing [13]. Although the ETSI standard is better than the Wiener filter in Mel-frequency domain, the proposed approach performs the best.

	WER (%)
Wiener filter (freq. averaging: rectangular window)	10.22
Wiener filter (freq. averaging: triangular window)	10.35
Wiener filter in Mel-frequency domain	11.24
ETSI Advanced front-end	10.57
RASTA ($J=10^{-7}$)	12.43

Table 2. Comparison of different frequency averaging methods.

By running separate recognition tests on speech corrupted with stationary and nonstationary noises, we observe that the gain due to averaging is more pronounced in nonstationary noise conditions. The results are tabulated in Table 3.

	WER (%)	
	No averaging	Averaging
Car noise	10.83	9.97
Factory noise	10.89	10.01
Back ground speech	11.23	10.15
Back ground music	11.28	10.23

Table 3. Performance in stationary and non-stationary background noise.

Finally we check the effectiveness of the averaging on other noise suppression schemes. From Table 4 it can be noted that the averaging consistently improves the performance.

	WER (%)	
	No Averaging	Averaging
Spectral Subtraction	11.27	10.54
Ephraim-Malah approach	11.33	10.49
Wiener filter	11.17	10.22

Table 4. Comparison of different noise reduction techniques.

4. CONCLUSION

We showed that simple time-frequency averaging of noise suppression filter could provide significant gains in the performance of automatic speech recognizers under noisy conditions. The averaged filter is more robust to errors in the estimation of *apriori* SNR, especially in non-stationary noisy conditions.

5. ACKNOWLEDGEMENT

This work has partially been funded by the European Union under the integrated project TC-STAR-Technology and Corpora for Speech-to-Speech Translation (IST-2002-FP6-506738, http://www.tc-star.org).

6. REFERENCES

[1] Boll S., "*Suppression of acoustic noise in speech using spectral subtraction"*, Acoustics, Speech, and Signal Processing, IEEE Transactions on Volume 27, Issue 2, Apr 1979 Page(s):113 - 120.

[2] J. S. Lim and A. V. Oppenheim, "*Enhancement and bandwidth compression of noisy speech,*" Proc. IEEE, vol. 67, pp. 1586–1604, Dec. 1979.

[3] M. Berouti, R. Schwartz, and J. Makhoul, "*Enhancement of speech corrupted by acoustic noise,*" in Proc. IEEE ICASSP, Washington, DC, Apr. 1979, pp. 208–211.

[4] R. J. McAulay and M. L. Malpass, "*Speech enhancement using a soft decision noise suppression filter*," IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-28, pp. 137–145, Apr. 1980.

[5] Y. Ephraim and D. Malah, "*Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator,"* IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-32, pp. 1109–1121, Dec. 1984.

[6] P. Lockwood and J. Boudy, "*Experiments with a nonlinear spectral subtractor (NSS), hidden Markov models and projection, for robust recognition in cars*," *Speech Communication*, vol. 11, pp. 215–228, June 1992.

[7] Virag N**., "***Single channel speech enhancement based on masking properties of the human auditory system*." Speech and Audio Processing, IEEE Transactions on, Volume 7, Issue 2, March 1999 Page(s):126 – 137.

[8] Y. N. Cheng and D. O'Shaughnessy, "*Speech enhancement based conceptually on auditory evidence*," IEEE Trans. Signal rocessing, vol. 39, pp. 1943–1954, 1991.

[9] Ing Yann Soon and Soo Ngee Koh, "*Speech enhancement using 2-D Fourier transform***",** Speech and Audio Processing, IEEE Transactions on Volume 11, Issue 6, Nov. 2003, Page(s):717 – 724.

[10] A. Adami, et al. , *Qualcomm-ICSI-OGI Features for ASR,* ICSLP-2002, Denver, Colorado, USA.

[11] Viikki O., Bye D. and Laurila K., "*A Recursive Feature Vector Normalization Approach for Robust Speech Recognition in Noise*", Proceedings of the International Conference on Acoustic, Speech and Signal Processing, Seattle, WA, USA, 1998.

[12] ETSI ES 202 050 recommendation, 2002. *Speech processing, transmission and quality aspects (STQ); distributed speech recognition; advanced front-end feature extraction algorithm; compression algorithms.*

[13] H. Hermansky and N. Morgan, "*RASTA processing of speech*", IEEE Trans. on Speech and Audio Processing, Volume 2, Issue 4, Oct. 1994, pp.578 – 589.

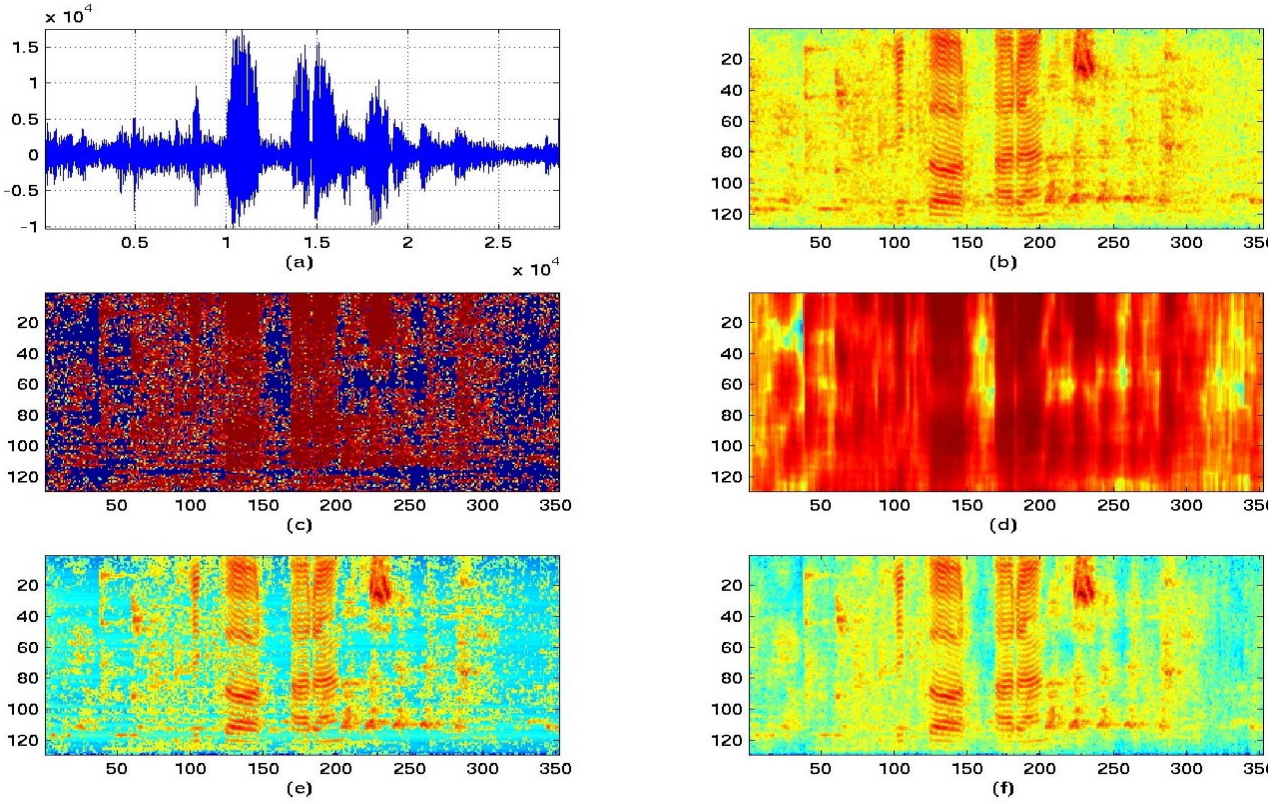

Figure 2. Speech spectrograms. (a) Noisy waveform (5dB SNR, background speech). (b) Spectrogram of noisy speech. (c) Gain of initial filter. (d) Gain of time-frequency smoothed filter. (e) Estimated clean spectrogram with initial filter. (f) Estimated clean spectrogram with smoothed filter.