SNR-DEPENDENT WAVEFORM PROCESSING FOR IMPROVING THE ROBUSTNESS OF ASR FRONT-END

Dušan Macho[*] *and Yan Ming Cheng*

Human Interface Lab, Motorola Labs
1301 E. Algonquin Road, Schaumburg, Illinois 90196, USA
email: dusan@gps.tsc.upc.es, ycheng@labs.mot.com

ABSTRACT

In this paper, we introduce a new concept in advancing the noise robustness of speech recognition front-end. The presented method, called SNR-dependent Waveform Processing (SWP), exploits SNR variability within a speech period for enhancing the high SNR period portion and attenuating the low SNR period portion in the waveform time domain. In this way, the overall SNR of noisy speech is increased, and at the same time, the periodicity of voiced speech is enhanced. This approach differs significantly from the well-known speech enhancement techniques, which are mostly frequency domain based, and we use it in this work as a complementary technique to them. In tests with SWP, we present significant clean and noisy speech recognition performance gains using the AURORA 2 database and recognition system as defined by ETSI for the robust front-end standardization process. Moreover, the presented algorithm is very simple and it is attractive also in terms of computational load.

1. INTRODUCTION

As the Automatic Speech Recognition (ASR) technology becomes more and more appealing to wireless applications, applications conducted in automobile environments or hands-free communication, the noise robustness determines the usability of ASR systems in these applications. The performance of current ASR systems radically deteriorates when input speech is interfered by noise (in most cases, background noise or background speech). This fact reduces the success of an ASR system in real-world applications. So far, many researchers' interests have been attracted and a large research effort has been conducted in the field of robustness. A survey of activities and past achievements in this field can be found in [1][2].

In this contribution, we limit ourselves on improving the robustness of front-end only, although, the robustness in both, front-end and back-end parts of an ASR system should be considered for practical applications. In the front-end part of ASR system, some techniques were adopted from the area of speech enhancement. For example, Spectral Subtraction (SS) and Wiener Filtering (WF) were successfully used to reduce the effect of additive noise on ASR spectral parameters. Both, SS and WF, are based on the idea of estimating noise in the frequency domain and removing the estimated noise spectrum from the noisy speech spectrum. For noise estimation, one usually relies on speech/non-speech detector to select noise frames (or segments) and update the noise estimate. However, a reliable speech/noise detector can be practically very difficult to build, especially in the case of non-stationary noises or low SNR noisy conditions. Thus, the assumption of good speech/noise detector is the fundamental weakness of these techniques. Recently, methods that do not need explicit speech/noise information have been proposed and they have been reviewed in [3].

Despite these successes, the problem of noise robustness is still not satisfactorily solved and efforts are still needed for further improvements in the front-end part of ASR system. In this paper, we explore a time domain based method as a complementary approach to the spectrum based speech enhancement techniques. The basic assumption is the existence of predictable SNR variability in the waveform time domain of voiced speech due to the speech periodicity and relatively constant noise energy.

The paper is organized as follows. In the next section, we describe the proposed algorithm. Then, experiments with the Aurora 2 database are presented to gain more insights into this algorithm. Finally, some remarks and conclusions are provided.

2. SNR-DEPENDENT WAVEFORM PROCESSING ALGORITHM

2.1 Basic idea

Within the period of voiced speech waveform, the instant speech energy reaches the highest point at the glottal closing instant (due to the highest glottal excitation) and the high energy will be sustained during the interval of closed glottis (the vocal tract damp is minimal). Once is the glottis opened, the instant energy is radically damped. Therefore, as is well known, the speech waveform (and also the instant speech energy contour) exhibits periodically maxima and minima. On the contrary, the interference noise energy generated by outside sources is relatively constant within the speech period. Therefore, within the noisy speech period, SNR is variable. It is relatively high during closed glottis and relatively low during opened glottis. Actually, this SNR variability is observable as long as the interference noise intensity is not extremely high. If one can locate the high SNR period portion and increase its energy (or, vice versa, locate the low SNR period portion and decrease its energy), the overall SNR of given voiced speech segment is enhanced. A front-end based on the SNR-enhanced signal is expected to be more robust. This forms the basis of our presented

[*] Dušan Macho is PhD student at UPC, Barcelona, Spain.

method – SNR-dependent Waveform Processing (SWP) – that can be related to the speech periodicity enhancement methods collected, e.g., in [4].

2.2 Algorithm description

In ASR systems, a frame-by-frame signal analysis is used to obtain the time evolution of speech spectral envelope, which is further used to generate parameters representing speech (usually, cepstral coefficients and their derivatives). In this process, the proposed SWP algorithm is applied on signal waveform that has been preprocessed by a spectral domain based speech enhancement technique. In this work, we use Two-stage Mel-warped Wiener Filter (2MWF, [5]) for preprocessing.

In SWP, for each frame, a smoothed instant energy contour is first computed (see Figure 1). We use the Teager energy operator [6] to obtain the instant energy value at each sample. On one hand, for voiced sounds, this smoothed energy contour has quasi-periodic property and its period depends on the actual fundamental frequency. On the other hand, for unvoiced sounds and silence/noise signal portions, a relatively flatter and random contour can be observed. Next, peaks (or maxima) of the smoothed energy contour are located by a simple peak-picking strategy. A windowing function $w(n)$ is constructed for each frame in such a way that a rectangular unit window of width W is placed between each two adjacent maxima found within the frame (see Figure 1 where the $w(n)$ has been multiplied by a constant in order to be visible).

Figure 1(a) shows waveform within a clean speech voiced frame together with both the corresponding smoothed energy contour and the windowing function obtained from it. Figure 1(b) shows the same frame with the file SNR equal to 0dB. As it can be observed from both figures, the rectangular windows are placed asymmetrically around each maximum since the high SNR portion (or the glottal close portion) in voiced frames is on the right side of each maximum. Prior to the mel-spectrum computation, the portions selected by windowing function are weighted more than the not selected – low SNR – portions. This operation, in fact, improves the SNR within voiced frames and enhances the signal periodicity. The original waveform within each frame, $s(n)$, is modified by using the windowing function $w(n)$ and a weighting parameter ε as follows

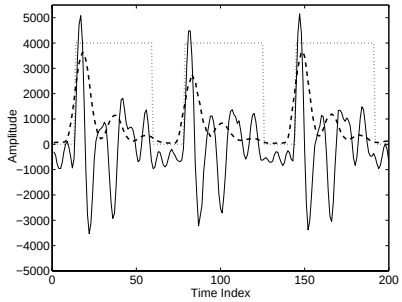
$$\begin{aligned} s_{SWP}(n) &= f(\varepsilon)\cdot s_{highSNR}(n) + \varepsilon \cdot s_{lowSNR}(n) \\ &= f(\varepsilon)\cdot w(n)s(n) + \varepsilon \cdot (1-w(n))s(n) \end{aligned}, \qquad (1)$$

where n is sample index and

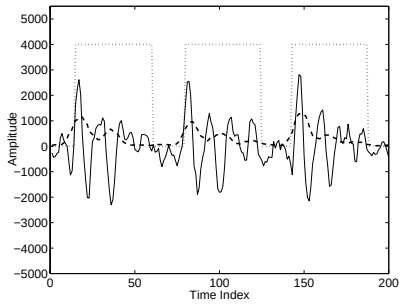
$$f(\varepsilon) = \sqrt{\frac{\sum_n |s(n)|^2 - \varepsilon^2 \cdot \sum_n |(1-w(n))s(n)|^2}{\sum_n |w(n)s(n)|^2}}, \qquad (2)$$

with $0 < \varepsilon \le 1$ and $f(\varepsilon) \ge 1$. The parameter ε determines the degree of attenuation of low SNR portions with respect to high SNR portions and $f(\varepsilon)$ is a function of ε that ensures the total frame energy after processing is the same as that before processing. Note that both W (the width of the rectangular function) and ε parameters must be experimentally determined.

An important advantage of the SWP is that it does not need a speech/non-speech detector. On the other hand, the fundamental weakness is that the interference noise energy should be sufficiently low to ensure correct maximum detection. However, as mentioned above, SWP is applied after 2MWF, which would have already enhanced the SNR to the adequate level.



(a)



(b)

Figure 1 Solid line in figure (a) shows typical clean speech waveform within a voiced frame. Smoothed energy contour (dashed line) and corresponding windowing function $w(n)$ (dotted line) are also shown. Figure (b) shows a low SNR version (SNR= 0dB) of the same frame as in figure (a). Note that in both cases, the 2-stage mel-warped Wiener filter preprocessing has been applied.

3. EXPERIMETAL RESULTS

3.1 Database, training-testing conditions and baseline performances

The AURORA database [7] is a noisy speech database distributed by the ETSI committee for the purpose of defining distributed speech recognition front-end standard. The database consists of clean and noisy connected digits. Both clean and noisy speech parts were prepared by filtering the TI database (both training and testing parts) using a telephone bandwidth filter (or filters). Additionally, to generate the noisy part, four different noises, such as exhibition hall, babble, suburban train and moving car noises, were artificially added to the clean TI data at various SNR levels (20dB, 15dB, 10dB, 5dB, 0dB and –5dB). AURORA 2 database is a re-release of the AURORA database with additive noises more representative in terms of real-world environments than its predecessor. Also, the

AURORA 2 database introduces additional mismatch conditions. There are two training scenarios: *multi-condition training* (MCT) and *clean speech training* (CST). In MCT, multiple noise types and SNR levels are used in training. On the other hand, in the CST scenario, only the clean speech portion of database is involved in training. Furthermore, within each of the training scenarios, three kinds of testing are performed: in test A, the training and testing data are matched in both channel effect and noise types; in test B, the training and testing data are only matched in channel effect, but not noise types; and finally, in test C, the channel mismatch between training and testing data is introduced.

The AURORA standardization committee provides a training script using HTK software in order to control the model topology, mixture specification and training process (generation of seed model, number of iterations, etc.). HMMs used here are whole word models for each digit and silence. The recognition grammar is the unknown length digit string. The committee also provides the baseline performance obtained by using 12 MFCCs + log energy coefficient and their delta and acceleration coefficients. In our subsequent experiments, we also use 12 MFCCs with log energy coefficient and their delta and acceleration coefficients appended, however, generated by a noise robust front-end.

3.2 Experiments on SWP

Two-stage mel-warped Wiener filtering [5] is used as baseline for SWP, because SWP is used after it as a complementary technique. Table 1 contains relative error reduction percentages with respect to the standard mel-cepstrum front-end performance for both the 2MWF technique alone and the 2MWF+SWP combination. The SWP technique has two parameters to be experimentally set: the width W of the high SNR frame portions and the low SNR portion attenuation factor ε. The rates in Table 1 have been obtained by using reasonable values for both parameters.

The best results were obtained with W=0.8 (in other words, the rectangular window width is 80% of the pitch period) and ε=0.8. Note that with these parameter values, the recognition performance in each condition is considerably improved with respect to the 2MWF baseline. Furthermore, the improvements in the clean speech training case are higher than those in the multi-condition training case.

From the previous tests we can deduce that the SNR improvement achieved within voiced frames by applying SWP leads to an improvement in noise robustness of the front-end. Actually, SWP improves also the periodicity of voiced speech – the fact that not only improves noise robustness but also increases the contrast between voiced and unvoiced speech that may help in clean speech recognition. In order to verify this hypothesis, we performed the above recognition tests with clean data. Error reduction percentages from these tests are shown in Table 2. Evidently, the SWP technique has doubled the performance improvement in comparison to the baseline 2MWF algorithm and, thus, the periodicity improvement may lead also to the clean speech recognition performance gain.

Additionally to the SWP idea, we thought that a frequency dependent SWP might further improve the robustness of front-end. In other words, we divided the speech frequency range into two bands (low and high frequency band) and we performed SWP separately for each band with different ε parameter value. However, we did not observe any further improvement. Since an insufficient number of experiments were conducted, we are not at the point to speculate any conclusion for the frequency dependent SWP approach.

In previous experiments, the width parameter W of the windowing function was kept constant across all speech utterances, once it was determined. However, it can be argued that for the high overall SNR utterances, one can select larger W in order to avoid speech distortion introduced by windowing. On the other hand, for low SNR utterances, a smaller W can be employed to achieve more aggressive noise reduction. For these purposes, we have modified the SWP algorithm in the way that W varies according to the frame SNR. The interval of W

Technique and parameter set	**Multi-Condition Training**			**Clean Speech Training**		
	A	**B**	**C**	**A**	**B**	**C**
2MWF (baseline)	26.37	21.54	33.81	47.03	53.76	37.04
2MWF+SWP, W=0.8, ε=0.9	27.71	24.57	35.09	50.86	55.43	43.86
2MWF+SWP, W=0.8, ε=0.8	***29.18***	***25.15***	***35.38***	***52.16***	***55.11***	***45.89***
2MWF+SWP, W=0.8, ε=0.7	26.62	23.47	33.89	52.92	54.94	47.17
2MWF+SWP, W=0.5, ε=0.9	27.78	25.20	34.52	50.72	55.59	44.16
2MWF+SWP, W=0.5, ε=0.8	27.81	24.95	33.64	51.81	55.52	46.13

Table 1 Relative error reduction percentages by using the 2MWF technique (baseline) and the combination of 2MWF and SWP in comparison to the MFCC standard in the AURORA 2 database.

Technique and parameter set	**Multi-Condition Training**	**Clean Speech Training**
	Clean Speech	**Clean Speech**
2MWF (baseline)	17.23	6.38
2MWF+SWP, W=0.8, ε=0.8	***32.43***	***13.01***

Table 2 Clean speech relative error reduction percentages by employing 2MWF and 2MWF+SWP in AURORA 2.

variation was set between 0.5 and 0.8, i.e., when the frame SNR is low, the rectangular window width is close to 50% of the pitch period and when the frame SNR is large, the window width is close to 80% of the pitch period. The frame SNR estimation is based on the difference between energies at the input and the output of 2MWF preprocessing. Table 3 lists error reduction percentages with SNR dependent *W*.

We can observe slight degradation in multi-condition training performances with *W* SNR-dependent, and significantly better performances for clean speech training with *W* SNR-dependent, as well. These results indicate that variable window width *W* does have some merit.

As a curiosity, we added an additional spectral subtraction (SS) algorithm [8] after SWP. For noise estimation, we used the waveform portions of each frame that have been indicated by SWP as low SNR. Table 4 shows that a further improvement can be obtained by additional SS in clean speech training tests. However, in the multi-condition training tests, considerable degradation in performances is observed in the used database.

Technique and parameter set	Multi-Condition Training			Clean Speech Training		
	A	B	C	A	B	C
2MWF, baseline	26.37	21.54	33.81	47.03	53.76	37.04
2MWF+SWP, ε=0.8, W=0.8	29.18	25.15	35.38	52.16	55.11	45.89
2MWF+SWP, ε=0.8, W_{SNR}=0.5-0.8	28.86	24.93	34.37	***54.34***	***57.18***	***46.77***

Table 3 Relative error reduction percentages by employing variable window width *W* in SWP (the last row).

Technique and parameter set	Multi-Condition Training			Clean Speech Training		
	A	B	C	A	B	C
2MWF, baseline	26.37	21.54	33.81	47.03	53.76	37.04
2MWF+SWP, ε=0.8, W=0.8	29.18	25.15	35.38	52.16	55.11	45.89
2MWF+SWP, ε=0.8, W_{SNR}=0.5-0.8	28.86	24.93	34.37	54.34	57.18	46.77
2MWF+SWP+SS, ε=0.8, W_{SNR}=0.5-0.8	27.26	23.78	33.25	***55.98***	***58.82***	***47.80***

Table 4 Relative error reduction percentages by using 2MWF+SWP and an additional spectral subtraction (the last row).

4. CONCLUSION AND REMARKS

In this paper, we introduced a new approach, SNR-dependent Waveform Processing (SWP), to improve the robustness of speech recognition front-end as well as speech recognition performance in general. The proposed method is based on the speech waveform processing and speech periodicity enhancement according to the instant SNR contour. It can be considered as a complementary technology to the well-known robust technologies, which are usually based on the frequency domain information.

SWP needs no speech/non-speech detector, which can be considered as fundamental weakness of many well-known technologies. However, it requires that the instant SNR contour in voiced sounds has a visible contrast, i.e., the interference noise should be relatively small. Due to this fact, we used SWP after speech enhancement (two-stage mel-warped Wiener filter).

We showed that by using SWP, both the robustness of existing front-end and the clean speech performance could be significantly improved in the AURORA noisy front-end standardization scenario. Finally, the small computation load of SWP makes this method even more attractive.

5. REFERENCES

[1] J.C. Junqua, J.-P. Haton, *Robustness in Automatic Speech Recognition*, Kluwer, 1996.

[2] Y. Gong, "Speech Recognition in Noisy Environment" Speech Communication, Vol. 16, 1995, pp. 261-291.

[3] S. Dupont, C. Ris, "Assessing Local Noise Level Estimation Methods", Workshop on Robust methods for speech recognition in adverse conditions, Tampere 1999, pp. 115-118.

[4] *Speech Enhancement*, Editor: J. Lim, Prentice-Hall, Englewood Cliffs, N.J., 1983.

[5] A. Agarwal, Y.M. Cheng, "Two-Stage Mel-Warped Wiener Filter for Robust Speech Recognition", ASRU Keystones, December 1999.

[6] H.M. Teager, "Some Observations on Oral Air Flow During Phonation," IEEE Trans. on Speech and Audio Processing, October 1980.

[7] D. Pearce, "Experimental Framework for the Evaluation and Verification of Distributed Speech Recognition Front-Ends," Version 4, Aurora ETSI, June 1998.

[8] P. Lockwood and J. Boudy, "Experiments with a Nonlinear Spectral Subtractor (NSS), Hidden Markov Models and the Projection for Robust Speech Recognition in Cars," Speech Communication, Vol. 11, 1992, pp. 215-228.