

Environment Adaptive Control of Noise Reduction Parameters for Improved Robustness of ASR

Chng Chin Soon¹, Bernt Andrassy², Josef Bauer²,
Günther Ruske¹

¹Department for Human-Machine-Interface, Technical University of Munich, Germany

zsz25@gmx.net

²Siemens AG, Corporate Technology IC5, Otto-Hahn-Ring 6, 81730 Munich, Germany

{bernt.andrassy, josef.bauer}@siemens.com

Abstract

This paper describes an extension to an automatic speech recognition system that improves the robustness concerning varying environments. A dedicated control unit tries to derive an optimal set of parameters for a Wiener Filter based noise reduction unit aiming at maximum recognition performances in different environments. The input measure for the control unit is derived from the speech signal. Apart from the SNR level, several other measures are investigated. The controlled parameters are closely related to the strength of the noise reduction. Several non-linear methods such as the Tabulated References and Neural Networks serve as the core of the control unit. Experiments on realistic handsfree as well as non-handsfree speech data show that the word error rate can be reduced by as much as 31% through the proposed methods. An already optimized static configuration of the applied noise reduction hereby serves as the baseline level.

1. Introduction

Noise reduction methods are essential to reduce the mismatch between the training data of an Automatic Speech Recognition system (ASR) and the application data in noisy conditions. In this paper, the Wiener Filter of the Siemens speech recognizer is investigated. Although the Wiener filter has already been optimized, observations have shown that while the recognition performance on noisy signals after Wiener filtering is generally improved, the performance on clean speech signals may be worsened through the filtering. To address this problem, two methods are presented. One introduces what we call the Signal to Noise Ratio (SNR) Threshold, which is used to determine for each signal whether the Wiener Filter should be activated or not. The other method attempts to estimate the Over-Estimation Factor (OEF) that will provide the optimum noise reduction. The OEF is closely linked to the intensity of the noise reduction from the Wiener Filter and hence reduces the effect of signal distortions. For the estimation of the optimal OEF, a Tabulated Reference, a Mahalanobis Classifier and a Neural Network were employed. These methods incorporate adaptivity to the Wiener Filter by updating its filter parameters during recognition.

In Section 2, the Threshold method is introduced. This is followed by the description of the OEF adaptation methods in Section 3. The experiments and results will be presented in Section 4.

2. Adaptive Control Using The Threshold Method

2.1. The SNR Threshold

The SNR Threshold Method makes use of the SNR of each utterance to determine whether the Wiener Filter should be applied. The SNR is calculated as the ratio of the energy of the noisy signal and the noise over the whole utterance taking into account the whole frequency range. The distinction between noisy signal and noise only parts is done with the help from either labels or a Voice Activity Detector (VAD). While using a VAD suggests a real life scenario, using the labels will provide an upper limit to the recognition improvement when an ideal VAD is employed. With an SNR Threshold, we implement a simple hard-decision rule which states

- If the SNR of the utterance is **LOWER** than the preset SNR Threshold, the signal **WILL** undergo noise reduction from the Wiener Filter during the preprocessing phase of the ASR.
- If the SNR of the utterance is **HIGHER** than the preset SNR Threshold, the signal **WILL NOT** undergo noise reduction from the Wiener Filter.

In other words, the Wiener Filter is applied only to utterances with relatively low SNR. To determine the preset SNR Threshold for our decision rule, all possible values of the SNR Threshold were investigated on the training data using the decision rule stated above. The threshold value that produced the least Word Error Rate (WER) on the training data was deemed as the optimal value and was used subsequently for our tests.

The left graph in Figure 1 illustrates the results of our training phase using SNR calculated with the labels. It shows the WER plotted against the range of SNR Threshold values that were investigated. The SNR values from our evaluation and training data lie within the range of -4dB to 38dB. The WER corresponding to the smallest SNR Threshold value of -5dB is the WER obtained with the Wiener Filter always off, since all the utterances have SNR values higher than this threshold. For the case of the WER corresponding to the largest SNR Threshold value on the graph, all utterances will be Wiener filtered (WF always on) as the SNR from all utterances are now lower than the threshold. This corresponds to the case of the baseline system described under Section 4. It can be seen from the graph that there is a minimum point at SNR Threshold 9dB with the minimum point from both the training and evaluation data coinciding with each other. This indicates that the optimum thresh-

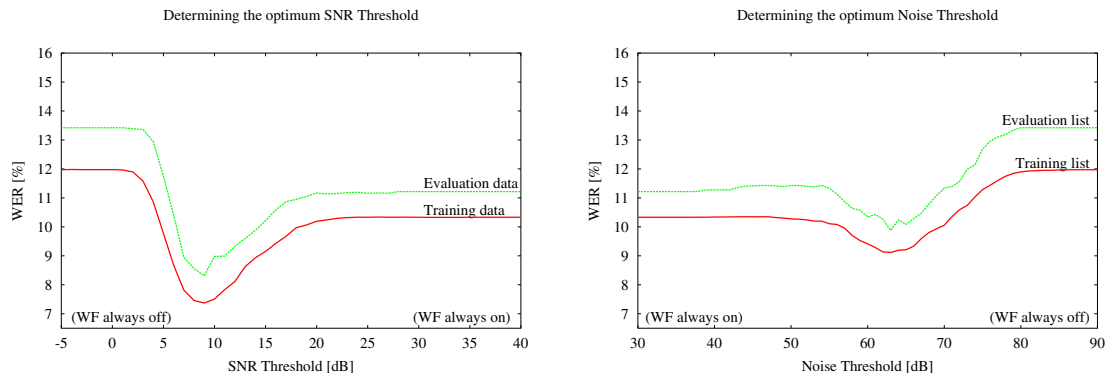


Figure 1: Recognition performance using SNR (left) and N (right) Threshold Methods on training and evaluation data.

old value determined from the training data, when used on the evaluation data, will also produce the lowest WER.

2.2. The N Threshold

We have extended the Threshold method to using the Noise Energy (N) as the deciding factor instead of the SNR in Section 2.1. The N of an utterance was taken as the average frame energy from the first four frames of the utterance. N has the principal advantage over the SNR that less effort is required to estimate it. Assuming correlation between N and SNR, N will provide us with a good indicator of the SNR and this method will behave similarly to the SNR Threshold Method with a slight decrease in performance.

Due to the inverse relationship between N and SNR, the hard-decision rule that we impose this time is the opposite of the SNR Threshold method. As seen from the right graph in Figure 1, a minimum point at 63dB is also obtained from the training data and it is also lower than the two WER values on the extreme ends. The WER value for the left end of the graph is the ASR performance when the Wiener Filter is applied on all utterances, regardless of the N of the utterances, and vice-versa for the right end. This optimum N Threshold value obtained from training also gives the lowest WER when applied on the evaluation data.

3. Adaptive Control of the Over-Estimation Factor

A problem concerning the Wiener Filter is the introduction of signal distortion in the form of musical tones. These distortions, as was observed in [1], arise due to the difference between the noise estimation, which is necessary for the calculation of the Wiener Filter coefficients, and the instantaneous noise energy. The estimation of the noise produces a smoothing effect and short-term fluctuations of the instantaneous noise around this estimated value will not be considered. This discrepancy between the estimated and the instantaneous noise causes residual isolated narrow spectral bands in the frequency domain after the noise reduction. These narrow bands produce the so-called musical tones which are disruptive for speech recognition, often resulting in higher error rates.

3.1. Effect of the OEF on Musical Tones

One common method to suppress musical tones is to introduce an Over-Estimation Factor (OEF) [2]. From the basic Wiener Filter equation

$$W(f) = \frac{P_{xx}(f)}{P_{xx}(f) + P_{nn}(f)} \quad (1)$$

where $P_{xx}(f)$ and $P_{nn}(f)$ are the power spectral densities (PSD) of the clean speech signal and noise respectively, we have extended it to incorporate the OEF as shown in Equation (2). The factor 2^α overestimates, or underestimates, the estimated noise PSD in order to prevent discrepancies due to the fluctuations of the instantaneous noise from causing the musical tones. In this paper, the term OEF will now refer to the α from the above-mentioned variable. For our experiments, it lies within the range of -4 to 12.

$$W(f) = \frac{P_{xx}(f)}{P_{xx}(f) + 2^\alpha P_{nn}(f)} \quad (2)$$

Current methods for estimating the OEF include assuming a linear relationship between the OEF and the SNR of the utterance [3], or a linear relationship with the pre-calculated estimates of the Wiener Filter coefficients [4]. Our investigations have shown, however, that there is a more complex relationship involved between the OEF and the various signal features and therefore a non-linear approach has to be taken to estimate the OEF accurately.

In this paper, we are using a classification approach to classify or map the OEF using certain features. To handle the classification task, we investigated three different methods, namely a Tabulated Reference, a Mahalanobis Classifier and a Neural Network.

3.2. Tabulated Reference

The idea of the tabulated reference is to map different SNR values to a corresponding optimal OEF. To determine the optimal OEF for the different SNR values, we first grouped all the utterances from the training data according to their SNR. For each SNR, we investigated all the possible OEF values within our above-mentioned range on all the utterances together within that group. The OEF that led to the lowest WER for the speech recognizer is picked out to be the optimal OEF for that SNR and is recorded in the reference. This process is repeated for all SNR

values. As can be seen from Table 1, which shows the range of SNR from 1dB to 10dB, there is no linear relationship concerning the OEF and the SNR.

SNR [dB]	opt. OEF	SNR [dB]	opt. OEF
1	2	6	5
2	4	7	5
3	3	8	5
4	4	9	1
5	5	10	0

Table 1: Tabulated Reference for SNR values in the range of 1dB to 10dB and their corresponding optimal OEFs.

This method has the advantage of having low complexity, a boon for the training and implementation stages. Apart from using the SNR, we have also repeated this method using N for reasons stated under Section 2.2. The results of both implementations can be seen under Section 4.2.

3.3. Mahalanobis Classifier

Another classification method for the OEF is to use a distance classifier. Following the idea that the OEF has a non-linear relationship to the signal features, each OEF is represented by several feature patterns distributed with a particular mean and variance within the feature space, i.e. each OEF is seen as a cluster within the N-dimensional space with N being the feature dimension. As features, we have used the SNR, SNR in sub-bands and N in sub-bands respectively (see Section 4). In this paper, we have picked the Mahalanobis Classifier for its ability to take into consideration the variances of the cluster distributions.

For the training phase, the optimal OEF for each utterance in the training data had to be determined. This was done by applying noise reduction on each utterance using all the possible values of the OEF and then testing them with the ASR. The OEF that produced the lowest WER is therefore the utterance’s optimal OEF. The process was repeated for every utterance in the training data and the results were used to train the Mahalanobis Classifier.

3.4. Neural Networks

We investigated a third method using Neural Networks [5] to handle the classification task. In our experiments a multilayer Perceptron Neural Network with error backpropagation was used. It consisted of three layers: input, hidden and output layer. For our experiments, the SNR sub-bands and N sub-bands described under Section 4 were used as input of the Neural Network. Output of the network was the optimal OEF obtained using the method described in Section 3.3.

3.5. Ambiguity of the OEF

One of the problems that we encountered while using these classification methods was that the optimal OEF calculated for training using the methods described above was ambiguous. We found that 95.0% of the utterances have more than one OEF that is optimal for it. Therefore, it was not possible to clearly map every utterance to only one optimal OEF each. To avoid this, for the Tabulated Reference, we simply picked the OEF with the smallest value for training. For the Mahalanobis classifier

and the Neural Network we investigated the following three approaches:

- Every optimal OEF of an utterance was used in training. That is, one utterance could be used several times with different target values during training.
- Out of all of the optimal OEFs determined from our training data, the most frequent one was taken and utterances that have this OEF as one of their optimal OEFs were assigned this value for training. From the remaining unassigned utterances, a new most frequent OEF was determined and the whole process was repeated until the fifth most frequent OEF was chosen and their corresponding utterances assigned. Only a few utterances were left unassigned and these were discarded.
- For every utterance the distance scores with each of the OEFs were calculated. The distance score was taken as the difference between the log probability of the forced recognition and the next best hypothesis. The OEF with the best distance score was recorded as the optimal OEF for that utterance.

4. Experiments

The training and evaluation data in our experiments consist of entries from the German SpeechDat II mobile database and also the German SpeechDat car database [6]. The mobile database includes recordings from a mobile phone in different environments such as office, car, public places, home, etc. The car database consists of recordings done in cars with a handsfree microphone. The training data consists of 4538 utterances from 1410 speakers while the evaluation data set has 867 utterances from 341 speakers. All utterances comprise of digit strings and they have been labeled, i.e. the exact frames for each digit as well as silence pauses are known. The labels are used to determine the SNR and the N of the utterances. This mimics the case of an ideal VAD and is used to portray the best case scenario.

The baseline system refers to the recognition results of the ASR when none of the proposed methods are applied. For this case, all utterances in the evaluation data are filtered using the Wiener Filter with the exact same filter parameters regardless of signal features. The OEF is set at a fixed value of 4 which is a value optimized for our speech recognizer.

For the experiments, different signal features were used. Other than the global SNR described in Section 2.1 and the global N from Section 2.2, there are two more features of interest:

- N sub-bands: Here N is calculated in 15 sub-bands. The sub-bands are determined by the bandwidth of the 15 mel-filters, used in our speech recognizer. In each sub-band, for the best case scenario, N is the frame energy averaged over all silence frames. For the simulation of a real life scenario, N will be the average energy of the first four frames of an utterance.
- SNR sub-bands: This is the SNR of the utterance calculated in 5 sub-bands. The intervals for each of these 5 sub-bands consist of 3 consecutive mel-scaled sub-bands. A resolution of 15 sub-bands may result in too little information available in each sub-band while a very low resolution will not take advantage of the frequency-selective characteristics of the noise. The SNR at each band can either be determined from labels or from a VAD (real life scenario).

4.1. Experiments using the Threshold Method

The methods investigated here are the SNR threshold and N threshold of the utterance. As seen from Table 2, it is clear that the results from the real life scenario (i.e. SNR calculation from VAD and N from the first 4 speech frames) are worse compared to the case with labels. As shown, the SNR Threshold Method is better compared to the N Threshold Method with an improvement of 26% over the baseline system for the ideal case.

The Threshold Methods		
Baseline	11.2%	
Features	SNR	N
Labels (Ideal VAD)	8.3%	9.9%
Real Life Scenario	(9.6%)	(10.0%)

Table 2: Recognition Performance of SNR and N Threshold Method.

Although the N Threshold led to worse results than the SNR-Threshold, it may still be a viable alternative as N is less complicated to estimate than the SNR.

4.2. Experiments using Adaptation of OEFs

Table 3 shows the results of the experiments done with the three methods to control the OEF of the Wiener Filter. Ambiguous OEFs were used for the training of the Neural Net and the Mahalanobis Classifier. The values in brackets show the results for the so called real-life scenario described above, otherwise they are results of the ideal VAD. We can see that for the Tabulated Reference Method again the SNR is a better feature than N. The experiments were done with different SNR and N resolutions of 1dB intervals and 5dB intervals and the results have shown to be slightly better for the higher resolution.

The Mahalanobis Classifier did not perform as well as the other two classification methods. Even the SNR Threshold Method has performed better. However, it is still able to achieve an improvement of 24% over the baseline system. This result is obtained when the SNR sub-bands are used as features, hinting at the potential of the sub-band approach.

The Neural Network performed the best among all methods. From the results shown, it was able to bring a 31% improvement over the baseline system, compared to 29% improvement from the Tabulated Reference, which has the next best results. Also, if you look at the values from the simulated real life scenarios which are shown in brackets in Tables 2 and 3, the Neural Network clearly shows the best results. Yet it comes with a price as it is comparatively expensive as well as complicated to implement. The Tabulated Reference is much simpler, saving time and effort.

Results from the methods mentioned in Section 3.5, dealing with the ambiguity of the OEFs, are shown in Table 4. The experiments were done using the feature N sub-bands determined under the real life scenario. For the Neural Network as well as the Mahalanobis Classifier, using the most frequent OEFs as optimal OEFs leads to the best results.

5. Conclusion

Two methods, each trying to solve the problem of a Wiener Filter regarding its environment dependent performance, have been introduced in this paper. One method utilizes an SNR

Baseline	11.2%		
Tabulated Reference			
Features	SNR		N
1dB Interval	8.0%	(9.0%)	8.9% (9.1%)
5dB Interval	8.1%	(9.6%)	8.9% (9.1%)
Mahalanobis Classifier			
Features	SNR	SNR sub-bands	N sub-bands
WER	8.9%	8.5%	9.7%
	(9.8%)	(9.2%)	(10.5%)
Neural Network			
Features	SNR sub-bands		N sub-bands
WER	8.1%	(8.9%)	7.7% (8.5%)

Table 3: Recognition performance of the different OEF adaptation methods.

	Mahalanobis	Neural Network
Ambig. opt. OEF	(10.5%)	(8.5%)
Most freq. OEF	(9.3%)	(8.3%)
Distance Scores	(10.5%)	(10.0%)

Table 4: Recognition performance of methods dealing with the OEF ambiguity.

Threshold to switch on and off the noise reduction by the Wiener Filter. The second method attempts to estimate the Over-Estimation Factor that will control the intensity of the noise attenuation during recognition. This method employs a Tabulated Reference, a Neural Network or a Mahalanobis classifier. It is shown that both methods bring about an improvement to the recognition performance with the latter producing better results. Furthermore the Neural Network was shown to be the most effective way to estimate the Over-Estimation Factor with a recognition improvement of 31% over the baseline system.

6. References

- [1] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," in *IEEE Trans. on Acoustics, Speech and Signal Processing*, no. 2, April 1979, pp. 113–120.
- [2] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *Proc. ICASSP*, 1979, pp. 208–211.
- [3] V. Schless and F. Class, "Snr-dependant flooring and noise overestimation for joint application of spectral subtraction and model combination," in *Proc. ICSLP*, 1998, pp. 1495–1498.
- [4] H. Puder, "Single channel noise reduction using time-frequency dependant voice activity detection," in *Proc. IWAENC*, 1999, pp. 68–71.
- [5] C. M. Bishop, *Neural Networks for Pattern Recognition*. Clarendon Press, 1998.
- [6] "Internet webpage of speechdat," 2003. [Online]. Available: <http://www.speechdat.org>