QUANTILE BASED NOISE ESTIMATION FOR SPECTRAL SUBTRACTION AND WIENER FILTERING

Volker Stahl, Alexander Fischer and Rolf Bippus

Philips Research Laboratories
Weisshausstrasse 2, D-52066 Aachen, Germany
email: {vstahl,afischer,bippus}@pfa.research.philips.com

ABSTRACT

Elimination of additive noise from a speech signal is a fundamental problem in audio signal processing. In this paper we restrict our considerations to the case where only a single microphone recording of the noisy signal is available. The algorithms which we investigate proceed in two steps: First, the noise power spectrum is estimated. A method based on temporal quantiles in the power spectral domain is proposed and compared with pause detection and recursive averaging. The second step is to eliminate the estimated noise from the observed signal by spectral subtraction or Wiener filtering. The database used in the experiments comprises 6034 utterances of German digits and digit strings by 770 speakers in 10 different cars. Without noise reduction, we obtain an error rate of 11.7%. Quantile based noise estimation and Wiener filtering reduce the error rate to 8.6%. Similar improvements are achieved in an experiment with artificial, non-stationary noise.

1. INTRODUCTION

The error rate of speech recognition systems increases dramatically in the presence of noise. It is therefore inevitable to provide some means of noise reduction in the front end of speech recognizers which operate under adverse conditions. A particularly noisy but important application domain of speech recognition is the car environment [3, 4, 5, 2, 8, 9]. In this paper we investigate different noise reduction methods and carry out experiments on a large speech database which has been recorded in the car.

The paper is structured as follows: In Section 2 we give a brief description of the speech recognition system and the database used for the experiments. Model assumptions on the speech and noise signal are stated in Section 3. In Section 4 we discuss two methods to estimate the noise power spectrum. The first method is based on frame wise speech / non-speech classification and recursive averaging over non-speech frames. As pause detection in noisy environments is a difficult problem, we propose a second method, which does not depend on a classifier. The noise is estimated as a temporal quantile in the power spectral domain. According to an experimental comparison, quantile based noise estimation performs significantly better, especially under non-stationary noise. In Section 5 we apply spectral subtraction and Wiener filtering to eliminate the estimated noise from the input signal. The results are summarized in Section 6.

2. DATABASE AND SPEECH RECOGNITION SYSTEM

The experimental results reported in this paper are based on the German digit string subset of the MoTiV database [7]. The corpus comprises 6034 utterances (4436 for training and 1598 for evaluating the error rate) by 770 speakers in 10 cars at various driving situations. Training and evaluation is always done on the matched scenario, i.e. the same noise elimination methods are applied during training and evaluation.

The speech recognizer is a continuous mixture density hidden Markov model (HMM) system whose parameters are estimated by Viterbi training. Each mixture consists of 8 Gaussian densities with density specific, diagonal covariance matrices. The system uses two HMMs for each digit, one for male and one for female speakers. The signal analysis is as follows: The observed speech signal is subdivided into overlapping, 16 ms spaced frames of 32 ms length. For each frame the power spectrum is estimated through a Hamming windowed FFT followed by a filter bank with 15 mel spaced triangular kernels. After a discrete cosine transform of the logarithmic filterbank outputs we obtain 12 mel frequency cepstral coefficients, which, augmented by 12 regression coefficients, are passed to the recognizer. In this paper we experiment with an additional preprocessing step in the power spectral domain in order to reduce additive noise in the signal.

3. NOTATION AND ASSUMPTIONS

We assume that the observed noise signal is a realization of a wide sense stationary process [11]. The major part of this paper deals with the estimation of its power spectrum $N(\omega)$. As the estimation is more reliable if more data is available, we use the notation $N(\omega, t)$ to denote an estimation of $N(\omega)$ using all frames from the beginning of the utterance up to frame t. Further, we assume that the clean speech signal within each frame t is an instance of a wide sense stationary process with power spectrum $S(\omega, t)$. For the sake of notational simplicity we do not distinguish between power spectra and periodigram based power spectrum estimations. As the speech and noise signal are assumed to be additive and independent, the power spectrum of the observed signal is $X(\omega, t) = S(\omega, t) + N(\omega)$. The power spectrum $X(\omega, t)$ is estimated by magnitude squared

Fourier coefficients of the observed signal in frame t. The clean speech signal power spectrum can therefore be estimated as $S(\omega,t) = X(\omega,t) - N(\omega,t)$.

4. ESTIMATION OF THE NOISE SPECTRUM

A crucial step in noise suppression methods like Wiener filtering or spectral subtraction is the estimation of the noise spectrum. There are applications where this task is simplified by some prior knowledge of the noise spectrum or by multi channel recordings. However, in this paper we assume that there is only a single microphone and all we know about the noise signal is that it is more or less stationary, independent of the speech signal and additive.

A commonly used method for noise spectrum estimation is to average over sections in the input signal which do not contain speech (Section 4.1). However, this approach requires that non-speech sections can be detected reliably, which is difficult especially under noisy conditions. Moreover, it relies on the fact that there actually exists a sufficient amount of non-speech in the signal. In order to avoid these problems, we propose a method to estimate the noise spectrum without explicit frame wise speech / non-speech classification (Section 4.2). The idea is to estimate the noise energy in each frequency band by temporal quantiles in the power spectral domain.

4.1. Noise Spectrum Estimation Based on Frame Wise Speech / Non-Speech Classification

If the signal to noise ratio is not too low, a simple method to detect speech is based on the signal energy. As the noise signal is assumed to be stationary, the signal energy in the entire utterance is greater or equal the noise energy. If the energy in a frame is significantly larger than the estimated noise energy, then the frame is likely to contain speech. Otherwise it is a pure noise frame and is used to update the current noise estimation. Let $X(\omega,t)$ be the power spectrum at frequency ω in the t-th frame of the input signal and $N(\omega,t)$ be the power spectrum of the estimated noise energy at frequency ω in frame t. A simple recursive formula to estimate the noise energy $N(\omega,t)$ is as follows:

$$N(\omega,t) = \begin{cases} N(\omega,t-1) & \text{if } \mathrm{XNR}(t) > \alpha \\ (1-\beta)N(\omega,t-1) + \beta X(\omega,t) & \text{else} \end{cases} \quad (1)$$

$$\mathrm{XNR}(t) = \frac{\sum_\omega X(\omega,t)}{\sum_\omega N(\omega,t-1)}$$

for all ω. The recursion is initialized by $N(\omega,0) = X(\omega,0)$, which reflects the assumption that the first frame of an utterance does not contain speech. Note that each frame is classified as either pure noise or speech plus noise. Equation (1) has two parameters α and β which depend on the speech data under consideration. Parameter α is related to the signal to noise ratio. Parameter β determines the adaptation speed of the noise estimation. According to experimental results $\alpha = 1.8$ and $\beta = 0.03$ perform well for the MoTiV corpus. The estimated noise $N(\omega,t)$ is removed from the input signal $X(\omega,t)$ by means of a Wiener filter, see Section 5. With this noise elimination method we obtain a word error rate of 10.3%. Without noise elimination the word error rate is 11.7%, i.e. the relative improvement is 12%.

Frame wise speech / non–speech classification under noisy conditions is a difficult problem far from being solved satisfactorily. The frame error rate of the speech / non-speech classifier described above is around 16% on the MoTiV corpus. In the next section we describe a method for estimating the noise spectrum which does not require explicit speech / non-speech classification.

4.2. Quantile Based Noise Spectrum Estimation

In [10] an algorithm for noise estimation based on minimum statistics has been proposed. As the minimum is sensitive to outliers we use a quantile different from minimum. The algorithm proposed in this section is somewhat simpler and has fewer parameters than the one in [10] but is computationally more expensive. A similar method has been described in [2].

It is well known that even in speech sections of the input signal not all frequency bands are permanently occupied with speech. In fact, a significant percentage of the time the energy in each frequency band is on the noise level. This observation can be used to estimate a noise power spectrum $N(\omega)$ from the observed speech signal $X(\omega,t)$ by taking the q-th quantile over time in every frequency band. More precisely, for every ω the frames of the entire utterance $X(\omega,t)$, $t = 0,\ldots,T$ are sorted such that $X(\omega,t_0) \le X(\omega,t_1) \le \ldots \le X(\omega,t_T)$. The q-quantile noise estimation is defined as

$$N(\omega) = X(\omega, t_{\lfloor qT \rfloor}). \quad (2)$$

For example, $q = 0$ yields the minimum, $q = 1$ the maximum and $q = 0.5$ the median. This approach is based on the assumption that each frequency band carries at least the q-th fraction of time only noise, even during speech sections. Obviously this is true for very small values of q but in order to obtain a robust estimation of the noise spectrum, which is not sensitive to outliers, we hope that q is somewhere near the median, i.e. $q \approx 0.5$.

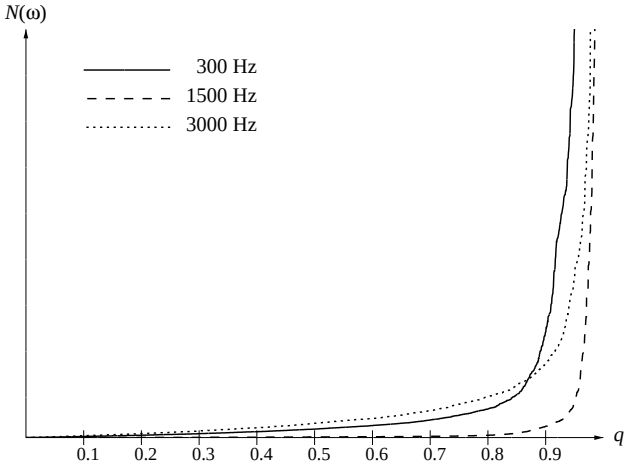

Figure 1: Quantiles of the energy distribution in the observed signal X at 300Hz, 1500Hz and 3000Hz for a typical utterance of the MoTiV corpus.

Figure 1 shows $N(\omega)$ according to (2) in dependence of q for 3 different frequencies ω and a typical 7 digit utterance taken from the MoTiV corpus. Roughly in 80-90% of the frames the signal energy in the frequency bands is low, i.e. close to the noise energy level and only in 10-20% of the time the frequency band carries high energy, voiced speech. Note that the curves also depend on the duration of the pause sections in the signal. However, the major part of the utterance in Figure 1 was speech. For the MoTiV corpus the optimal value for q was determined experimentally. The estimated noise $N(\omega)$ was eliminated from the signal by a Wiener filter, see Section 5. The resulting word error rates (WER) are summarized in Table 1. The word error rate without any noise reduction method is 11.7%, i.e. the relative reduction is 26% for the optimal choice $q = 0.55$. With a 5909 words test set we obtain under certain simplifying assumptions a confidence interval of 0.8% on the 95% significance level for the baseline error rate 11.7%. Error rates below 10.9% are therefore significant improvements.

q	0.2	0.3	0.4	0.5	0.55	0.6	0.7
WER	11.3	10.8	10.1	8.9	8.6	8.8	9.7

Table 1: Word error rate with Wiener filter and noise estimation $N(\omega)$ according to (2).

Causality. Note that the estimation of the noise spectrum depends on the entire utterance $X(\omega, t)$ for all $t = 0, \ldots, T$. A noise suppression filter based on this approach is therefore not causal. However, if we define $N(\omega, t)$ as the q-quantile of $X(\omega, \tau)$ for $\tau = 0, \ldots, t$, we obtain a causal filter. Table 2 summarizes the results of the same experiments as in Table 1 but this time we used a causal noise estimation. The error rates achieved by the causal filter are slightly higher than for the non-causal case. The reason is that the noise estimation at the beginning of the signal is very unreliable because few data is available to estimate $N(\omega, t)$ for small t.

q	0.2	0.3	0.4	0.5	0.55	0.6	0.7
WER	11.5	10.8	10.0	8.8	8.9	9.1	10.2

Table 2: Same experiment as in Table 1 but with causal noise estimation.

Efficiency. The computational cost and memory consumption for estimating $N(\omega, t)$ grows with t. This is problematic for real time and low resource implementations. As a consequence we investigated approximate methods for the quantile computation which are more efficient in terms of time and space. The idea is to store the observations $X(\omega, t)$ for $t = 0, 1, \ldots$ in a buffer with fixed length Δ. Separate buffers are used for each frequency ω. If a buffer is full, then the largest and the smallest element are removed from the buffer. The quantile is determined by considering only the elements in the buffer. The obvious question now is how large the buffer should be and how much the recognition error rate increases with a finite length buffer. Results of experiments with different buffer lengths Δ and $q = 0.5$ are reported in Table 3. As expected, the error rate increases for small buffer sizes and achieves asymptotically the error rate of the exact quantile computation. Another method to improve efficiency is to integrate several adjacent frequencies and do a band wise noise estimation [6].

Δ	3	5	10	20	40	60	100
WER	10.6	10.2	9.3	9.1	9.3	9.2	8.9

Table 3: Same experiment as in Table 2 for $q = 0.5$ but with limited buffer length Δ for the quantile computation.

Non-stationary Noise. We observed that the classifier based method in Section 4.1 performs quite poorly if the noise energy increases abruptly, say at time $\hat{t}$. The reason is that the estimated noise $N(\omega, \hat{t})$ at time $\hat{t}$ is small compared to subsequent input frames $X(\omega, t)$ for $t > \hat{t}$, especially if frame $X(\omega, t)$ does not contain speech. Therefore, according to (1), all frames after $\hat{t}$ are classified as speech and hence the noise estimation will not be updated any more after time $\hat{t}$, i.e. $N(\omega, t) = N(\omega, \hat{t})$ for all $t > \hat{t}$. In other words, the noise estimation does not converge to the observed noise. The quantile based method presented in this section does not suffer from this problem and seems therefore advantageous for non-stationary noise. In order to verify this theoretical consideration by an experiment, we inserted 0.5 seconds of car noise from a BMW 540 at 50 km/h before the beginning of each sound file of the test set. The columns of Table 4 contain the word error rates for the cases no noise reduction, noise estimation by the classifier based method and noise estimation by the quantile based method for $q = 0.5$ and buffer sizes 10, 20, 60, and unlimited respectively. In each scenario the error rate is significantly higher than in the corresponding case without inserted car noise. The deterioration for the classifier based noise estimation method, however, is much more severe than for the quantile based method and is even worse than for the case without noise elimination. The adaptation time to a changing noise signal in the quantile based method is proportional to the buffer length Δ, which explains why in this experiment shorter buffer lengths give better results.

Method	none	classifier	quantile $\Delta = 10, 20, 60, \infty$			
WER	13.7	18.5	10.1	10.5	10.6	11.7

Table 4: Word error rate if 0.5 seconds low energy car noise are added to the beginning of the sound files of the test set.

5. ELIMINATION OF THE NOISE FROM THE SPEECH SIGNAL

In the previous section we discussed methods for estimating the noise power spectrum $N(\omega, t)$. In this section we review approaches for eliminating the estimated noise from the observed signal. If we had complete information about the noise spectrum, i.e. magnitude and phase, the noise elimination would amount to a simple subtraction of the complex Fourier coefficients. Unfortunately we have no phase information of the noise. Hence we apply spectral subtraction and Wiener filtering for the noise elimination. The FIR Wiener filter is defined as the linear filter which minimizes the mean square error in the time domain. Spectral subtraction relies on the fact that the power spectrum of the sum of two independent random signals is the sum of the power spectra. The noise elimination rule of spectral

subtraction is therefore simply to subtract the power spectrum of the estimated noise from the power spectrum of the observed signal. Surprisingly the formulae for the Wiener filter and spectral subtraction are quite similar. Let

$$H(\omega,t) = (X(\omega,t) - N(\omega,t)) \ / \ X(\omega,t). \qquad (3)$$

The noise reduced signal $S(\omega,t)$ by Wiener filtering is

$$S(\omega,t) \quad = \quad H(\omega,t)^2 X(\omega,t),$$

noise reduction by spectral subtraction is defined as

$$S(\omega,t) = X(\omega,t) - N(\omega,t) = H(\omega,t)X(\omega,t).$$

Sometimes the long term estimated noise power spectrum $N(\omega,t)$ can be larger than the instantaneous observed power spectrum $X(\omega,t)$. In this case we would expect that the noise reduced power spectrum $S(\omega,t)$ should be zero. Therefore (3) is usually modified as

$$H(\omega,t) = \max(X(\omega,t) - N(\omega,t),\ 0) \ / \ X(\omega,t)$$

Experimental experience indicates that better recognition results are achieved if a small fraction of the noise power is left in the signal [1, 10]. Hence, the energy of the noise reduced signal $S(\omega,t)$ which is passed to the recognizer is

$$S_\gamma(\omega,t) = \max(\ S(\omega,t),\ \gamma N(\omega,t)\ )$$

where $\gamma = 0.04$ has been chosen experimentally.

An experimental comparison of spectral subtraction and Wiener filtering for $\gamma = 0.04$ is given in Table 5. The noise power spectrum $N(\omega,t)$ has been estimated as in Table 2.

q	0.2	0.3	0.4	0.5	0.6	0.7	0.8
Wiener	11.5	10.8	10.0	8.8	9.1	10.2	12.1
Subtr.	11.7	11.4	10.9	10.1	9.9	9.6	11.6

Table 5: Experimental comparison of the word error rates of Wiener filtering and spectral subtraction.

As suggested in [1] the performance of spectral subtraction can be improved by subtracting an overestimation of the noise power spectrum, i.e.

$$S_{\gamma,\eta}(\omega,t) = \max(X(\omega,t) - \eta N(\omega,t),\ \gamma N(\omega,t)).$$

In our experiments we found an optimum for $\eta = 2.5$, which gives a word error rate of 9.2% for $q = 0.5$.

6. CONCLUSION

We investigated methods to remove additive noise from a speech signal which has been recorded in the car environment by a single microphone. The error rate of a speech recognizer has been reduced by up to 26% relative by quantile based noise estimation in the power spectral domain and Wiener filtering. The methods proceed in two steps: Estimation of the noise signal and elimination.

Noise Estimation. We studied two noise estimation methods: The first one is based on frame wise speech/non-speech classification and recursive smoothing over non-speech frames (Section 4.1), the second method estimates the noise by quantiles in the power spectral domain (Section 4.2). The quantile based noise estimation method gives significantly better results but is more expensive in terms of computing time and memory. An approximation algorithm for improving the efficiency of the quantile based method has been proposed. The classifier based method requires prior knowledge about the signal to noise ratio, which is not the case for the quantile based method. However, the quantile based method relies on assumptions on energy distributions of human speech in the time–frequency domain, which need to be verified by more experiments. Finally, the quantile based method seems to work better for certain kinds of non-stationary noise than the classifier based method.

Noise Elimination. Two methods for removing the estimated noise have been investigated, namely spectral subtraction and Wiener filtering. The latter seems superior according to experimental evidence (Section 5). If spectral subtraction is modified such that an appropriate overestimation of the noise is subtracted, then the achieved error rate comes close to the Wiener filter.

7. REFERENCES

[1] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of Speech Corrupted by Acoustic Noise," in *Proc. ICASSP*, (Washington, USA), pp. 208–211, Apr. 1979.

[2] H. G. Hirsch and C. Ehrlicher, "Noise Estimation Techniques for Robust Speech Recognition," in *Proc. ICASSP*, pp. 153–157, 1995.

[3] Juang, B. H. "Speech Recognition in Adverse Environments", Computer Speech and Language 5: pp. 275-294, 1991.

[4] Junqua, J.-C., Haton, J.P. "Robustness in Automatic Speech Recognition: Fundamentals and Applications", Kluwer, Boston, 1996.

[5] 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, pp. 215–228, 1992.

[6] L. Singh and S. Sridharan, "Speech Enhancement using Critical Band Spectral Subtraction," in *Proc. ICSLP*, (Sydney, Australia), Nov. 1998.

[7] D. Langmann, T. Schneider, R. Grudszus, A. Fischer, T. Crull, H. Pfitzinger, M. Westphal, and U. Jekosch, "CSDC - The MoTiV Car-Speech Data Collection," in *First International Conference on Language Resources and Evaluation*, (Granada, Spain), May 1998.

[8] A. Fischer and V. Stahl, "Subword Unit based Speech Recognition in Car Environments," in *Proc. ICASSP*, (Seattle, USA), pp. 257–261, May 1998

[9] A. Fischer and V. Stahl, "Database and Online Adaptation for improved Speech Recognition in Car Environments," in *Proc. ICASSP*, (Phoenix, USA), pp. 445–449, March 1999

[10] R. Martin, "Spectral Subtraction based on Minimum Statistics," *Proc. European Signal Processing Conference*, pp. 1182–1185, Sep 1994.

[11] M. H. Hayes, "Statistical Digital Signal Processing and Modeling," *John Wiley & Sons, Inc.*, 1996.