NOISE ESTIMATION TECHNIQUES FOR ROBUST SPEECH RECOGNITION

H. G. Hirsch*, C. Ehrlicher

Institute of Communication Systems and Data Processing, Aachen University of Technology, 52056 Aachen, Germany

ABSTRACT

Two new techniques are presented to estimate the noise spectra or the noise characteristics for noisy speech signals. No explicit speech pause detection is required. Past noisy segments of just about 400 ms duration are needed for the estimation. Thus the algorithm is able to quickly adapt to slowly varying noise levels or slowly changing noise spectra. This techniques can be combined with a nonlinear spectral subtraction scheme. The ability can be shown to enhance noisy speech and to improve the performance of speech recognition systems. Another application is the realization of a robust voice activity detection.

1. INTRODUCTION

Many proposals are known to improve speech recognition in situations with a noisy background, e.g. [1], [2], [3], [4]. Especially the modified statistics of spectral parameters should be considered in case of using HMMs [5]. A well working algorithm to detect speech pauses is presumed to determine these modified statistics.

This contribution presents two methods to estimate the spectral parameters of noise without an explicit speech pause detection. The first algorithm calculates the noise level in each subband as a weighted average of past spectral magnitude values which are below an adaptive threshold. The second approach evaluates the histograms of past spectral magnitude values in each subband. The maximum is taken as an estimate for the noise level.

2. ESTIMATION OF NOISE SPECTRUM

Most of the noise reduction techniques based on single channel recordings need an estimation of the noise spectrum. This is usually done by detection of speech pauses to evaluate segments of pure noise. In practical situations this is a difficult task especially if the background noise is not stationary or the signal-to-noise ratio (SNR) is low. Some approaches are known to avoid the problem of speech pause detection and to estimate the noise characteristics just from a past segment of noisy speech [3], [6], [7]. The disadvantage of most approaches is the need of relatively long past segments of noisy speech.

The first method presented here calculates the weighted sum of past spectral magnitude values X_i in each subband i. The weighting is done by a simple first order recursive system

$$\hat{N}_i(k) = (1-\alpha) * X_i(k) + \alpha * \hat{N}_i(k-1) \qquad (1),$$

where $X_i(k)$ denotes the spectral magnitude at time k in subband i and $N_i(k)$ is an estimation of the noise magnitude.

Some algorithms immediately use an average of past spectral power values as an estimation for the noise power in the individual subband to realize a so called continous spectral subtraction (CSS) [1]. In contrast to these approaches an adaptive threshold is introduced here. The magnitude values X_i are distributed according to a Rayleigh distribution in segments of pure noise. Considerably higher values occur at the onset of speech. Thus a threshold $\beta * \hat{N}_i(k-1)$ is introduced where β takes a value in the range of about 1,5 to 2,5. When the actual spectral component $X_i(k)$ exceeds this threshold this is considered as a rough detection of speech and the recursive accumulation is stopped. The accumulated value is taken as an estimation for the noise level at this time. This simple processing is illustrated in figure 1 as part of a complete noise reduction scheme.

The noise estimate $\hat{N}_i$ is calculated with a first order recursive system. $\hat{N}_i$ is multiplied with an over-estimation factor β in the usual range of about 1,5 to 2,5. For positive values of $(X_i - \beta \hat{N}_i)$ the data input as well as the recursive accumulation are stopped. This indicates an onset of speech. Negative values of $(X_i - \beta \hat{N}_i)$ are set to zero to get an estimate S_i of the

*) This author is with Ascom Business Systems (Solothurn, Switzerland) now. Email: hirsch@ens.ascom.ch

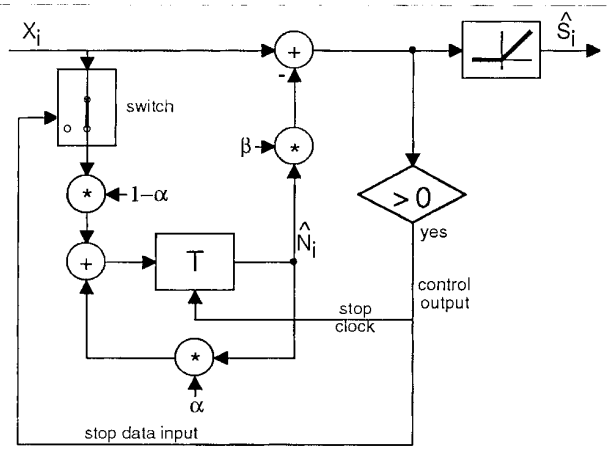

Fig. 1. Simple noise reduction scheme in one subband

clean speech. The computational complexity is low.

The second approach is based on histograms of past spectral values in each subband. The above mentioned threshold is used to evaluate histograms of past values which are below this threshold. This processing can be interpreted as a rough separation of the distributions of noise (Rayleigh distribution) and speech. Speech takes much higher values. Past values corresponding to noise segments of about 400 ms duration are evaluated to determine the distribution in about 40 bins. The noise level is estimated as maximum of the distribution in each subband. The estimated values for the noise magnitude are smoothed versus time to eliminate rarely occuring spikes. This leads to a very accurate estimation of the noise spectrum.

An objective evaluation of the accuracy is illustrated in figure 2. Different stationary noise signals [8] were artificially added to clean speech at different SNRs. The average noise spectrum $\underline{N}$ is calculated from the noise itself as well as an average estimated noise spectrum $\hat{\underline{N}}$ obtained with the two mentioned techniques. The relative error

$$\frac{\sum_i (\hat{\underline{N}}_i - \underline{N}_i)^2}{\sum_i \underline{N}_i^2} \qquad (2)$$

is calculated as an objective measurement for the accuracy. In figure 2 the average relative error is shown adding a car noise [8] to different utterances of 3 male and 3 female speakers. Average spectral components are calculated as sum over all frames of a FFT based spectral analysis.

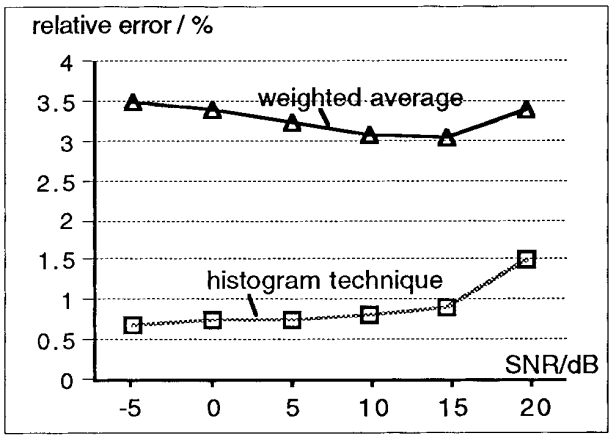

Fig. 2. Relative error for the estimation of the noise power spectrum with both techniques at different SNRs

A good estimation can be achieved by both techniques. As expected the evaluation of histograms leads to better results. The increase at higher SNR is caused by an inaccurate noise estimation during segments of speech. Even at a high SNR this small errors effect the calculation of a relative error much more.

Both techniques are applied to a noise reduction scheme using nonlinear spectral subtraction [2]. A well working suppression was confirmed by informal listening tests. Also negative effects as e.g. musical tones can be reduced by optimizing parameters, e.g. the overestimation factor.

3. RECOGNITION OF NOISEX DATA

A first series of recognition experiments was carried out using the isolated words of the Noisex92 study [8]. This is a first attempt for a common data base to get comparative results on the recognition of noisy speech. Different noises are artificially added to utterances of the ten digits at different SNRs. The digits were spoken 100 times separately for training and testing. Recordings exist for a male and a female speaker at a sampling rate of 16 kHz. Both above mentioned estimation techniques are applied to the nonlinear spectral subtraction as a preprocessing step to recognition. A HMM recognizer [9] is used for the experiments configured as a connected word recognizer. A single mixture continous HMM is trained

with 8 emitting states for each word. Pauses are represented by a single state HMM. All training is done with the clean data only. A set of 15 MFCC (Mel frequency cepstral coefficients) are calculated as acoustic parameters for the recognition. Some results are shown in figure 3 as average of 5 different noises and as average of the two speakers. The male and the female utterances were separately recognized.

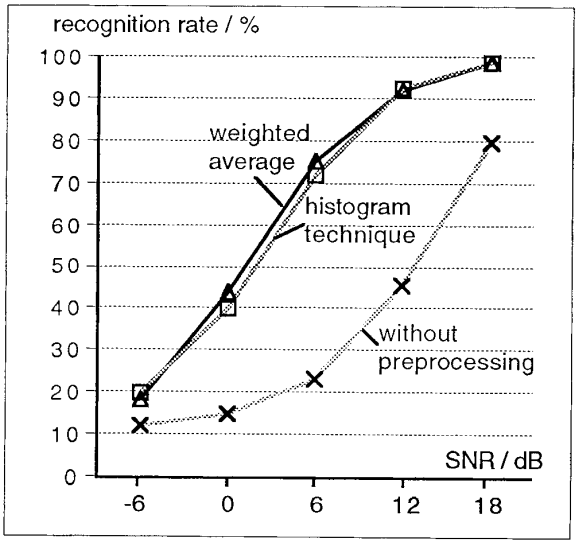

Fig. 3. Average recognition results for a speaker dependent recognition of the Noisex data

Considerable improvements can be achieved by applying the noise estimation techniques. In addition the detection of speech pauses is implemented to obtain these results. This is necessary because no individual HMM model is calculated for the pauses at each noise condition. The detection is based on the evaluation of the SNRs in all subbands. A relative measure NX_{rel} of the ratio N/X (noise to noise&signal) is calculated for each subband:

$$NX_{irel}(k) = \frac{NX_i(k) - NX_{imin}(k)}{NX_{imax}(k) - NX_{imin}(k)} \qquad (3)$$

where smoothed versions are used for N and X. NX_{min} and NX_{max} are determined from past segments of about 600 ms. The value NX_{rel} is already calculated to realize the nonlinear spectral subtraction. A low value of NX_{rel} indicates speech. Speech pauses are detected by counting the number of subbands where the ratio NX_{rel} is less than a certain threshold e.g. in our realization a value of 0,4. Using a FFT filter bank with 128 subbands frames are classified as pauses if the number of "active" bands is less than 4. A robust voice activity detection can be achieved by this technique.

During segments of pauses the spectral subtraction is applied with an overestimation factor of 3. An interesting result is observed decreasing the overestimation factor from usual values in the range of 2,5 to a value of just 1 for segments of speech. Best results are obtained for an overestimation factor in the range of 1. The use of a factor of 1 for the overestimation degrades the noise reduction scheme to a simple subtraction in subbands. This effect can be explained by the training of the HMMs. The average and the variance of the acoustic parameters are estimated for each state from the clean data. The modified increased variance of spectral parameters is not considered in this contribution. Thus the average values are mainly evaluated for the recognition. Subtracting more than the noise level will lower the spectral parameters in the individual subband on the average. This will decrease the estimated averages in the corresponding states of the HMM.

4. SPEAKER INDEPENDENT RECOGNITION

A second data base is considered for another series of experiments. 13 words (digits inluding "zero", "oh" and "yes", "no") were recorded from 200 speakers via telephone lines. This time a HMM recognizer is configured as an isolated word recognizer but including a model for the pauses. A continous HMM is trained with 8 emitting states and 4 mixtures per state. Pauses are represented by a single state with 4 mixtures. 5 PLP cepstral coefficients [10] are used as acoustic parameters. For each condition the recognition rate is calculated as an average of 4 recognition experiments using 50 different speakers out of the 200 for training and the remaining 150 for testing in each individual experiment. Car noise was artificially added at SNRs in the range from 5 to 20 dB.

Some recognition results are illustrated in figure 4. The experiments applying the simple noise reduction scheme shown in figure 1 were done in comparison to using PLP [10] or Rasta-PLP analysis [11]. Rasta-PLP is a well working technique to reduce the influence of different frequency responses during recording or transmission. It introduces a high-pass filtering of the logarithmic spectral envelopes in each subband. Thus it can be interpreted as a spectral subtraction in the logarithmic domaine. The impulse response of the high-pass filter is similiar to the response of the

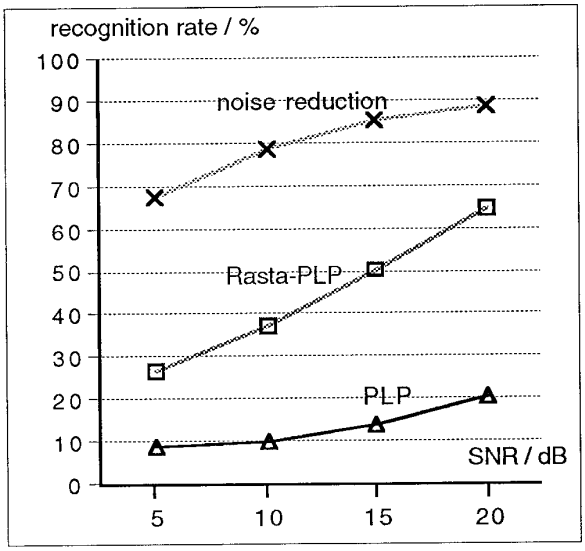

Fig. 4. Recognition results for a speaker independent recognition (car noise)

filter scheme presented in figure 1.
The simple noise reduction is integrated into PLP analysis. PLP includes a spectral analysis with a FFT and the calculation of a number of subband energies where the definition of the subbands is derived from the frequency groups of the human auditive system. Better results are obtained applying the noise reduction to the subband energies of the 15 nonlinearly spaced filters than to all output values of the FFT. The variance of the subband energy seems to decrease by summing up the FFT energies in 15 subbands during segments of noise.
Also this time a speech pause detection is added in case of applying the processing scheme to speech recognition. Thus the 15 subband energies are filtered with the mentioned filter scheme using an overestimation factor of 2. A robust speech detection can be realized summing up the output values of all filters and looking for a positive value of the sum. Again a filtering is applied with an overestimation factor of 3 during segments of noise. Speech segments are filtered with a factor of 1.

5. CONCLUSION

Two methods are presented to estimate the noise spectra and more general the noise characteristics of noisy speech without an explicit speech pause detection. These are able to adapt to varying noise levels. Also one of the algorithms has a low computational complexity. The approaches can be combined with well known spectral subtraction techniques. Reducing the overestimation factor to a value in the range of 1 leads to simple reduction schemes with low computational complexity.
These approaches are a good supplement to HMM recognition schemes which consider the modified statistics of spectral parameters caused by additive noise [5].

6. ACKNOWLEDGEMENT

This work was partly carried out at the International Computer Science Institute in Berkeley, USA. The authors would like to thank the whole speech group and especially Dr. Morgan for fruitful discussions and a stimulating athmosphere.

7. REFERENCES

[1] J.A. Nolazco Flores, S. J. Young, "Continous Speech Recognition in Noise Using Spectral Subtraction and HMM Adaptation", ICASSP-94, Vol.1, pp. 409-412, 1994
[2] P. Lockwood, J. Boudy, "Experiments with a Nonlinear Spectral Subtractor, Hidden Markov Models and the Projection for Robust Speech Recognition in Cars", Speech Communication, Vol. 11, No. 2-3, pp. 215-228, 1992
[3] D. Van Campernolle, "Noise Adaptation in a Hidden Markov Model Speech Recognition System", Computer Speech and Language, Vol. 3, pp. 151-167, 1989
[4] H.G. Hirsch, P. Meyer, H.W. Rühl, "Improved Speech Recognition Using High-Pass Filtering of Subband Envelopes", Eurospeech-91, pp. 413-416, 1991
[5] M.J.F. Gales, S.J. Young, "Cepstral Parameter Compensation for HMM Recognition in Noise", Speech Communication, Vol.12, No. 3, pp. 231-240, 1993
[6] R. Martin, "An Efficient Algorithm to Estimate the Instantaneous SNR of Speech Signals", Eurospeech-93, pp.1093-1096, 1993
[7] H.G. Hirsch, "Estimation of Noise Spectrum and its Application to SNR Estimation and Speech Enhancement", Technical Report TR-93-012, International Computer Science Institute, Berkeley, USA, 1993
[8] A. Varga, H.J.M. Steeneken, " Assessment for Automatic Speech Recognition: II. Noisex92: A Database and an Experiment to Study the Effect of Additive Noise on Speech Recognition Systems", Speech Communication, Vol.12, No. 3, pp. 247-252, 1993
[9] S.J. Young, "HTK Version 1.4: Reference Manual and User Manual", Cambridge University Engineering Department, Speech Group, 1992
[10] H. Hermansky, "Perceptual Linear Predictive (PLP) Analysis of Speech", JASA, pp. 1738-1752, 1990
[11] N. Morgan, H. Hermansky et al., "Compensation for the Effect of the Communication Channel in Auditory-Like Analysis of Speech (Rasta-PLP)", Eurospeech-91, pp. 1367-1370, 1991