Approaches to Robust Speech Recognition in Mobile Devices

Tim Fingscheidt, Panji Setiawan, Sorel Stan
Siemens AG, COM Mobile Devices, Grillparzerstr. 10-18, D – 81675 Munich, Germany
Email: {first name}.{last name}@siemens.com

Abstract

Automatic speech recognition in mobile devices has to cope with varying acoustical background noises in potentially low SNR situations. Therefore, techniques such as noise reduction are required to ensure the accuracy of the speech recognition process. This paper compares different kinds of environment compensation techniques, all operating frame-wise, either in the spectral or in the log-spectral domain. We report on word recognition rate as well as on word accuracy, the later also being a performance measure in absence of speech (i.e. only background noise) cases.
As a classical technique, we first investigate Wiener filtering using a voice-activity-driven noise power spectral density (psd) estimation. Then we perform a comparison with the more advanced recursive least squares (RLS) weighting rule for speech enhancement, as well as with the use of minimum statistics as noise psd estimation. Finally, simulation results with the S-IMM (sequential interacting multiple models) approach are shown. It turns out that approaches known from speech enhancement perform very well also for speech recognition. This allows the use of the specific noise reduction function in a mobile device for speech telephony on the one hand, and for robust speech recognition on the other hand.

1 Introduction

Approaches to robustness in automatic speech recognition cover noise reduction as part of speech enhancement, model-based, and data-driven environment compensation. They operate in the spectrum, mel, or log-mel domain. This paper focuses on those approaches well suited for applications in mobile devices with their tight constraints to complexity and memory. We are investigating Wiener filtering [1], recursive least squares (RLS) [2], and a model-based technique which employs Kalman filtering in the log-spectral domain, i.e., sequential interacting multiple models (S-IMM) [3].
The rest of the paper is organized as follows: The speech enhancement techniques are presented in the next section. Experimental results for speech recognition and some analysis on the algorithms are given in Section 3.

2 Speech Enhancement Techniques

In speech enhancement, the following assumptions are commonly used:

$$Y_k(m) = S_k(m) + N_k(m), \quad \hat{S}_k(m) = H_k(m) \cdot Y_k(m),$$

where $Y_k(m), S_k(m), N_k(m)$ and $H_k(m)$ denote the noisy speech, clean speech, and noise spectra, and a particular weighting rule, respectively, for a single frame m and frequency bin k.

2.1 Wiener Filtering

The weighting rule is derived based on the minimization of the mean squared error (MMSE) in the frequency domain. The *a-priori* SNR (signal to noise ratio) based weighting rule $H_k^W(m)$ is defined as:

$$H_k^W(m) = \frac{\xi_k(m)}{\xi_k(m)+1}, \tag{1}$$

where the *a-priori* SNR estimate $\xi_k(m)$ is updated using the decision-directed approach [4]:

$$\xi_k(m) = \max\{(1-\varepsilon)\cdot P[\vartheta_k(m)-1] + \varepsilon \cdot \frac{|\hat{S}_k(m-1)|^2}{\alpha \cdot \hat{\lambda}_{N_k}(m)}, 0.01\},$$

with $P[x] = \max\{x, 0\}$. The *a-posteriori* SNR estimate is calculated as:

$$\vartheta_k(m) = \frac{|Y_k(m)|^2}{\alpha \cdot \hat{\lambda}_{N_k}(m)},$$

with $\hat{\lambda}_{N_k}(m)$ being the noise power estimate and α being an overestimation factor. A good choice is $\varepsilon = 0.89$. Choosing however $\varepsilon = 0$, approximately the (well known) *a-posteriori* SNR based Wiener filtering follows.

2.2 Recursive Least Squares

This technique is originally derived based on the minimization of the weighted least squares error criterion yielding a weighting rule (we call it *a-posteriori* RLS) $H_k^{RLS}(m)$ given by [2]:

$$H_k^{RLS}(m) = \frac{E_{Y_k}(m)}{E_{Y_k}(m) + \alpha \cdot E_{N_k}(m)}, \tag{2}$$

where $E_{Y_k}(m) = \sum_{\mu=0}^{m} \rho_{Y_\mu} \cdot |Y_k(\mu)|^2$ and $E_{N_k}(m) = \sum_{\mu=0}^{m} \rho_{N_\mu} \cdot \hat{\lambda}_{N_k}(\mu)$. Assuming $\rho_{Y_\mu} = \lambda_Y^{m-\mu}$, $E_{Y_k}(m)$ is calculated recursively:

$$E_{Y_k}(m) = \lambda_Y \cdot E_{Y_k}(m-1) + |Y_k(m)|^2 \tag{3}$$

$$E_{N_k}(m) = \lambda_N \cdot E_{N_k}(m-1) + \hat{\lambda}_{N_k}(m). \tag{4}$$

Note that the use of two distinct coefficients, $\lambda_Y = 0.1$ and $\lambda_N = 0.05$, compensates for the use of $E_{Y_k}(m)$ instead of a theoretically more justified $E_{S_k}(m)$ in the weighting rule.
In a modified weighting rule the term $E_{Y_k}(m-1)$ in (3) can be replaced by

$$E_{\hat{S}_k}(m-1) = |\hat{S}_k(m-1)|^2 + \lambda_S \cdot E_{\hat{S}_k}(m-2) \tag{5}$$

with $\lambda_S = 0.1$, yielding the *a-priori* RLS weighting rule.

	Wiener		RLS		S-IMM
	post.	prior.	post.	prior.	
VAD	86.88%	89.13%	88.75%	88.99%	-
MS	89.45%	90.18%	88.40%	88.36%	-
Other	-	-	-	-	82.81%

Table 1: Performance in word accuracy.

	Wiener		RLS		S-IMM
	post.	prior.	post.	prior.	
VAD	87.93%	90.77%	88.80%	89.58%	-
MS	90.62%	91.54%	89.17%	89.15%	-
Other	-	-	-	-	83.57%

Table 2: Performance in word recognition rate.

2.3 Noise Estimation Techniques

The noise power estimate $\hat{\lambda}_{N_k}$ which occurs in the calculation of both weighting rules shown previously is computed by two different techniques. The first is the voice-activity-driven (VAD) technique which employs the speech/non-speech condition to update the current noise power estimate. If moderate speech activity is assumed, then the estimate of the noise psd is increased by:

$$\hat{\lambda}_{N_k}(m) = (1-\varepsilon_{up}) \cdot \hat{\lambda}_{N_k}(m-1) + \varepsilon_{up} \cdot \overline{|Y_k(m)|^2}, \quad (6)$$

and if no speech activity is assumed, it decreases

$$\hat{\lambda}_{N_k}(m) = (1-\varepsilon_{dn}) \cdot \overline{|Y_k(m)|^2} + \varepsilon_{dn} \cdot \hat{\lambda}_{N_k}(m-1). \quad (7)$$

The previous estimate $\hat{\lambda}_{N_k}(m-1)$ is taken if strong speech activity is assumed. The smoothed observation power is calculated as:

$$\overline{|Y_k(m)|^2} = (1-\varepsilon_Y) \cdot \overline{|Y_k(m-1)|^2} + \varepsilon_Y \cdot |Y_k(m)|^2. \quad (8)$$

The second noise estimation technique is the well-known minimum statistics (MS) [5]. This technique is basically tracking the minimum value of the smoothed power spectra within a finite window. This technique performs very well in speech enhancement and is implemented without any modifications.

3 Experimental Results

Aurora 3 German digits database has been used to evaluate the perfomance of the approaches. The corpus has been recorded in a real car environment and has three different training and test cases, i.e., well matched, medium mismatch, and high mismatch. The recognizer is running with 25/10 ms frame length/shift and taking 39 features as input, generated from a linear discriminant analysis (LDA) on 2 consecutive frames of 39 MFCCs.

The performance is measured in *word accuracy*, ACC $= (N-D-S-I)/N \times$ 100%, and *word recognition rate*, WRR $= (N-D-S)/N \times$ 100%. The symbols N, D, S, I denote the total number of reference words, number of deletion errors, substitution errors, and insertion errors, respectively.

Tables 1 and 2 show the performance. It becomes obvious that the speech enhancement approaches outperform the S-IMM. The model used in S-IMM does not perform well especially in mismatch conditions. Moreover, all RLS approaches are much better than classical (a-posteriori SNR based) Wiener filtering. The a-priori SNR based Wiener filtering however yields a better performance than RLS. Using minimum statistics (MS) in RLS does not bring advantages. A reason may be that the MS noise estimate is already quite flat over time, while the RLS has specific control (λ_N) to smooth the often times varying VAD based noise estimate. The a-priori type of Wiener filter with MS shows the best performance with more than 10% relative word error rate improvement over any RLS technique and more than 25% relative to standard Wiener filtering with a VAD based noise psd estimation.

Conclusions

In this paper we compared several approaches to robust speech recognition in mobile devices. Fortunately, the best scheme (a-priori SNR based Wiener filter with minimum statistics noise psd estimation) can be used in a mobile phone in synergy with speech enhancement for handsfree telephony. The scheme decreases the word error rate by more than 25% relative to standard Wiener filtering with a VAD based noise psd estimation. The a-priori type of RLS formulation gives also already good performance and is particularly interesting for low resource implementations. S-IMM has been found to perform quite poor on car noise.

Acknowledgement

The authors would like to thank S. Suhadi for his important contributions to this paper.

References

[1] P. Scalart and J. Vieira Filho, "Speech Enhancement Based on A Priori Signal to Noise Estimation," in *Proc. of ICASSP'96*, Atlanta, GA, pp. 629–632, May 1996.

[2] C. Beaugeant, V. Gilg, M. Schoenle, P. Jax, and R. Martin, "Computationally Efficient Speech Enhancement Using RLS and Psycho-acoustic Motivated Algorithm," in *Proc. of World Multi-conference on Systemics, Cybernetics and Informatics*, 2002.

[3] N. S. Kim, "Feature Domain Compensation of Nonstationary Noise for Robust Speech Recognition," in *Speech Communication*, vol. 37, pp. 231–248, 2002.

[4] Y. Ephraim and D. Malah, "Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator," in *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 32, no. 6, pp. 1109-1121, Dec. 1984.

[5] R. Martin, "Noise Power Spectral Density Estimation Based on Optimal Smoothing and Minimum Statistics," in *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 5, pp. 504-512, July 2001.