Robust Speech Recognition for Mobile Devices in Car Noise

Panji Setiawan, Suhadi Suhadi, Tim Fingscheidt, and Sorel Stan

Siemens AG, COM Mobile Devices
Grillparzerstr. 10-18, D – 81675 Munich, Germany
{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. Its performance in car noise environments is of our particular interest. We put focus on noise reduction techniques as applicable for speech enhancement to ensure the accuracy of the speech recognition process. We report on word recognition rate as well as on word accuracy, the latter also being a performance measure in the 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. Mel based root-cepstral coefficients has been taken as an alternative to the conventional Mel-frequency cepstral coefficients (MFCCs). The *a-priori* SNR based Wiener filtering with the minimum statistics and Mel based root-cepstral coefficients achieves 33.16% relative improvement in word accuracy over the classical technique.

1. Introduction

Approaches to robustness in automatic speech recognition cover noise reduction as part of speech enhancement, model-based, and data-driven environment compensation. This paper focuses on those approaches well suited for applications in mobile devices with their tight constraints to complexity and memory.

We are particularly interested in those approaches which have been developed in speech enhancement area. This is motivated by the fact that speech enhancement applications have been widely used in mobile devices whereas speech recognition applications have yet to cope with the problems in the presence of background noise.

Noise reduction techniques in speech enhancement are the focus of our investigation. Some of the techniques to be investigated in this paper are Wiener filtering [1] and recursive least-squares (RLS) [2]. The *a-priori* and *a-posteriori* definitions of the corresponding weighting rules have been extensively used in this paper. The noise power spectral density (psd) is estimated by means of minimum statistics [3] and a conventional voice-activity-driven techniques.

The root function as an alternative to the logarithmic function, which has been shown to increase robustness for speech enhancement and speech recognition in the presence of background noise [4–7], is implemented. This function was initially proposed in [8], which operates in the linear frequency domain.

AURORA 3 German digit corpus [9] has been taken to evaluate the system since it mostly fulfills our interest to focus on speaker independent digit dialling task in car noise environments.

The rest of the paper is organized as follows: noise reduction techniques are presented in the next section. Root function as an alternative to logarithmic operation is briefly discussed in Section 3. Experimental results for speech recognition and some analysis on the algorithms are given in Section 4.

2. Noise Reduction Techniques

Noise reduction techniques in speech enhancement are commonly using the following assumptions:

$$Y_k(m) = S_k(m) + N_k(m), \tag{1}$$

$$\hat{S}_k(m) = H_k(m) \cdot Y_k(m), \tag{2}$$

where complex variables $Y_k(m), S_k(m), N_k(m)$ and real variable $H_k(m)$ denote the noisy speech, clean speech, and noise spectra, and a particular weighting rule, respectively, for a frame m and frequency bin k. The processing also assumes statistical independency between frequency bins. It is generally assumed that phase information is of low importance in the design process of a weighting rule.

2.1. Wiener Filtering

The weighting rule is derived based on the minimization of the mean squared error (MMSE) in the frequency domain using the following assumption for the expectation function:

$$E\{S_k^*(m)N_k(m)\} = 0. \tag{3}$$

The *a-priori* SNR (signal to noise ratio) based weighting rule $H_k^W(m)$ is defined as [1]:

$$H_k^W(m) = \frac{\xi_k(m)}{\xi_k(m)+1}, \tag{4}$$

where the *a-priori* SNR estimate $\xi_k(m)$ is updated using the decision-directed approach [10]:

$$\xi_k(m) = (1-\varepsilon) \cdot P[\vartheta_k(m) - 1] + \varepsilon \cdot \Gamma_k(m-1), \tag{5}$$

with $0 \leq \varepsilon \leq 1$ and $P[x] = \max\{x, 0\}$. The term $\Gamma_k(m-1)$ is calculated as

$$\Gamma_k(m-1) = \frac{|\hat{S}_k(m-1)|^2}{\alpha \cdot \hat{\lambda}_{N_k}(m-1)}. \tag{6}$$

The *a-posteriori* SNR estimate is defined as:

$$\vartheta_k(m) = \frac{|Y_k(m)|^2}{\alpha \cdot \hat{\lambda}_{N_k}(m)}, \tag{7}$$

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.

In our implementation, the *a-priori* SNR estimate in (5) has a minimum value of 0.01 and $\hat{\lambda}_{N_k}(m-1)$ in (6) is replaced by $\hat{\lambda}_{N_k}(m)$.

2.2. Recursive Least-Squares Algorithm

This technique is originally derived based on the minimization of the weighted least-squares error criterion yielding a weighting rule:

$$H_k^{LS}(m) = \frac{E_{S_k}(m)}{E_{S_k}(m) + \alpha \cdot E_{N_k}(m)}, \tag{8}$$

with

$$E_{S_k}(m) = \sum_{\mu=0}^{m} w(\mu) \cdot |S_k(\mu)|^2 \tag{9}$$

$$E_{N_k}(m) = \sum_{\mu=0}^{m} w(\mu) \cdot \hat{\lambda}_{N_k}(\mu), \tag{10}$$

where $w(\mu)$ denotes a real function which assigns weights to current and previous frames. The following assumption has been used to obtain (8):

$$\sum_{\mu=0}^{m} \rho(\mu) \cdot S_k^*(\mu) N_k(\mu) = 0. \tag{11}$$

The variable $E_{S_k}(m)$ is not possible to compute, therefore we use the term *a-posteriori* RLS $H_k^{RLS}(m)$ to refer to the weighting rule given by [2]:

$$H_k^{RLS}(m) = \frac{E_{Y_k}(m)}{E_{Y_k}(m) + \alpha \cdot E_{N_k}(m)}, \tag{12}$$

where

$$E_{Y_k}(m) = \sum_{\mu=0}^{m} w_Y(\mu) \cdot |Y_k(\mu)|^2 \tag{13}$$

$$E_{N_k}(m) = \sum_{\mu=0}^{m} w_N(\mu) \cdot \hat{\lambda}_{N_k}(\mu). \tag{14}$$

Taking $w_Y(\mu) = \rho_Y^{m-\mu}$ and $w_N(\mu) = \rho_N^{m-\mu}$ for $0 \leq \rho_Y, \rho_N \leq 1$, the variables $E_{Y_k}(m)$ and $E_{N_k}(m)$ are calculated recursively:

$$E_{Y_k}(m) = \rho_Y \cdot E_{Y_k}(m-1) + |Y_k(m)|^2 \tag{15}$$

$$E_{N_k}(m) = \rho_N \cdot E_{N_k}(m-1) + \hat{\lambda}_{N_k}(m). \tag{16}$$

The term *recursive* has been added to merely indicate the recursive psd computation in (15) and (16). Note that the use of two distinct coefficients, $\rho_Y = 0.1$ and $\rho_N = 0.05$, 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 (15) can be replaced by

$$E_{\hat{S}_k}(m-1) = \rho_S \cdot E_{\hat{S}_k}(m-2) + |\hat{S}_k(m-1)|^2 \tag{17}$$

with $\rho_S = 0.1$, yielding what we call the *a-priori* RLS weighting rule. A floor value of 0.09 is used for the weighting rule $H_k^{RLS}(m)$ in (12).

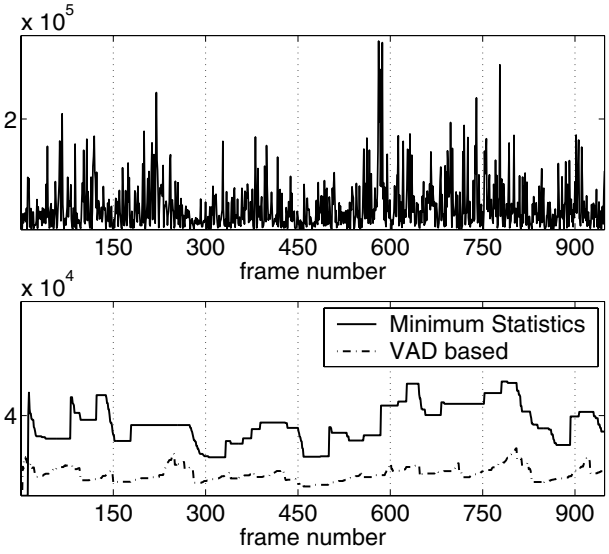

Figure 1: *Top: Noise psd at frequency 1156.25 Hz; Bottom: Noise psd estimates at frequency 1156.25 Hz.*

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:

$$\hat{\lambda}_{N_k}(m) = (1 - \varepsilon_{up}) \cdot \hat{\lambda}_{N_k}(m-1) + \varepsilon_{up} \cdot \overline{|Y_k(m)|^2}, \tag{18}$$

and if no speech activity is assumed, it decreases to

$$\hat{\lambda}_{N_k}(m) = (1 - \varepsilon_{dn}) \cdot \overline{|Y_k(m)|^2} + \varepsilon_{dn} \cdot \hat{\lambda}_{N_k}(m-1). \tag{19}$$

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

$$\overline{|Y_k(m)|^2} = (1 - \varepsilon_Y) \cdot \overline{|Y_k(m-1)|^2} + \varepsilon_Y \cdot |Y_k(m)|^2. \tag{20}$$

The second noise estimation technique is the well-known minimum statistics (MS) [3]. This technique is basically tracking the minimum value of the smoothed noisy power spectra within a finite window. This technique performs very well in speech enhancement and is implemented without any modifications. It is computationally more expensive than the first technique.

Figure 1 shows the psd of a park noise which is taken from NTT corpus and the noise psd estimates from both techniques at frequency 1156.25 Hz.

3. Root-Cepstral Coefficients

It has been indicated in [5] that the logarithmic deconvolution is not necessarily the optimal scheme for speech analysis. Several functions exist as an alternative to the log $(\cdot)$ operator, such as:

$$f(\cdot) = \begin{cases} \frac{1}{\gamma}\left[(\cdot)^\gamma - 1\right], & \gamma \neq 0 \\ \log(\cdot), & \gamma = 0, \end{cases} \tag{21}$$

which is known as the generalized logarithmic function [11], and a direct root function [8]:

$$f(\cdot) = (\cdot)^{\gamma}, \tag{22}$$

where γ is a real number with $-1 < \gamma < 1$.

It has been shown that an optimal value of γ has the following advantages over the logarithmic operation:

- better estimation of the pole-zero model of vocal tract impulse response which is represented by the full cepstral coefficients calculated on compressed spectral coefficients in linear frequency domain,
- more robust cepstral representation in the presence of background noise, especially for the high-order cepstral indexes,
- bigger effect resulting from the first-order pre-emphasis operation.

All of these lead to more robust representation of acoustic features in the cepstral domain. The improvements were also reported even when a noise reduction technique is used. Please note that initial development of the root function was in the linear frequency domain.

The use of Mel based root-cepstral coefficients for speech recognition, where the root function is applied after the Mel-frequency warping, was initially reported in [7] and showed significant robustness in car noise environments. Later experiments conducted on AURORA 2 corpus also showed improvements in several noisy conditions [4].

In this paper, the root function in (22) is used to compute the Mel based root-cepstral coefficients as an alternative to the Mel based log-cepstral coefficients widely known as the MFCCs.

4. Experimental Results

4.1. Database

AURORA 3 German digits database [9] has been used to evaluate the perfomance of the approaches. The corpus was recorded in a real car environment and has three different training and test cases, i.e., well matched, medium mismatch, and high mismatch. There are 2929 speech files in total which have been separately divided into training and testing sets. The files in each set have been distributed to represent each of the three cases.

4.2. Recognizer

A whole-word Hidden Markov Models (HMMs) based recognizer is used. The recognizer is running with 25/10 ms frame length/shift and taking 39 dimensional features as input, generated from a linear discriminant analysis (LDA) on 2 consecutive observation frames. Each observation frame is a 39 dimensional vector consisting of 12 MFCCs/root-cepstral coefficients and log energy, 13 delta, and 13 acceleration coefficients. The 12 cepstral coefficients are computed with a discrete cosine transform (DCT) operation on 15 log/root-compressed Mel filter bank outputs.

4.3. Performance measurement and analysis

The performance of the approaches is measured in *word accuracy*,

$$\text{ACC} = \frac{N - D - S - I}{N} \times 100\%, \tag{23}$$

Table 1: *Word accuracy performance of noise reduction + MFCCs techniques.*

	Wiener		RLS	
	post.	prior.	post.	prior.
VAD based	86.88%	89.13%	88.75%	88.99%
MS	89.45%	89.67%	88.40%	88.36%

Table 2: *Word accuracy performance of noise reduction + Mel based root-cepstral coefficients techniques.*

	Wiener+10th-root		RLS+10th-root	
	post.	prior.	post.	prior.
VAD based	87.87%	90.50%	88.78%	89.50%
MS	89.99%	91.23%	88.52%	88.85%

Table 3: *Relative word accuracy improvement of the Mel based root-cepstral coefficients (table 2) over the MFCCs (table 1) technique.*

	Wiener+10th-root		RLS+10th-root	
	post.	prior.	post.	prior.
VAD based	7.58%	12.60%	0.31%	4.63%
MS	5.12%	15.10%	0.99%	4.21%

and *word recognition rate*,

$$\text{WRR} = \frac{N - D - S}{N} \times 100\%. \tag{24}$$

The relative improvement of a certain performance measurement value p [%] over a reference value q [%] is calculated as

$$\text{relative [ACC / WRR]} = \frac{p - q}{100 - q} \times 100\%. \tag{25}$$

The symbols N, D, S, I denote the total number of reference words, number of deletion errors, substitution errors, and insertion errors, respectively. The overall weighted performance is computed with the following weights: 0.4, 0.35, and 0.25 for well matched, medium mismatch, and high mismatch, respectively.

Tables 1 and 4 show the performance of the noise reduction techniques with the standard MFCCs features in word accuracy and word recognition rate, respectively. Minimum statistics brings considerable improvement to classical (*a-posteriori* SNR based) Wiener filtering relative to the VAD based noise estimate, i.e. more than 19% in word accuracy. The best technique is the *a-priori* SNR based Wiener filtering with minimum statistics which gives at least 6% relative increase in word accuracy over any of the RLS techniques.

Moreover, all RLS approaches are much better than classical Wiener filtering. Using minimum statistics in RLS does not bring advantages. A reason may be that the minimum statistics noise estimate is already quite flat over time, while the RLS has specific control (ρ_N) to smooth the often times varying VAD based noise estimate.

Tables 2 and 5 show the performance in word accuracy and word recognition rate, respectively, if the Mel based root-cepstral coefficients are used instead of the MFCCs. The relative improvements of all noise reduction techniques with Mel based root-cepstral coefficients over the MFCCs are shown in tables 3 and 6. The root function increases the robustness of all techniques in word accuracy.

Table 4: *Word recognition rate performance of the noise reduction + MFCCs.*

	Wiener		RLS	
	post.	prior.	post.	prior.
VAD based	87.93%	90.77%	88.80%	89.58%
MS	90.62%	91.07%	89.17%	89.15%

Table 5: *Word recognition rate performance of the noise reduction + Mel based root-cepstral coefficients.*

	Wiener+10th-root		RLS+10th-root	
	post.	prior.	post.	prior.
VAD based	88.65%	91.62%	89.22%	90.18%
MS	90.94%	92.11%	89.05%	89.57%

Table 6: *Relative word recognition rate improvement of the Mel based root-cepstral coefficients (table 5) over the MFCCs (table 4) technique.*

	Wiener+10th-root		RLS+10th-root	
	post.	prior.	post.	prior.
VAD based	5.97%	9.21%	3.75%	5.80%
MS	3.41%	11.64%	-1.15%	3.92%

Improvements are significant for almost all types of Wiener filtering compared to RLS. *A-priori* SNR based Wiener filtering gains the biggest relative improvement in word accuracy, i.e. 15.10%. This implies that a relative improvement of 33.16% in word accuracy has been achieved compared to the classical Wiener filtering with VAD based noise estimate and MFCCs.

The performance of *a-priori* SNR based Wiener filtering with both the conventional VAD based noise estimate and minimum statistics is almost comparable when the Mel based root-cepstral coefficients are used. Considering a much higher complexity of the minimum statistics technique compared to the VAD based, a trade-off for a slightly worse performance with the VAD based noise estimation is afforded.

5. Conclusions

In this paper we compared several approaches to robust speech recognition in mobile devices. 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, which uses the Mel based root-cepstral coefficients, increases the word accuracy by more than 33% relative to classical Wiener filtering with a VAD based noise psd estimation which uses the MFCCs features. The *a-priori* type of RLS formulation, as well as the use of of VAD based noise estimate for *a-priori* type Wiener filtering give also already good performance and are particularly interesting for low resource implementations.

6. References

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

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

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

[4] Yapanel, U., Hansen, J.H.L., Sarikaya, R., and Pellom, B., "Robust Digit Recognition in Noise: An Evaluation Using the AURORA Corpus", EUROSPEECH 2001, Aalborg, Denmark, September 3-7, 2001.

[5] Alexandre, P. and Lockwood, P., "Root Cepstral Analysis: A Unified View. Application to Speech Processing in Car Noise Environments", Speech Communication, vol. 12, pp. 277-288, 1993.

[6] Alexandre, P., Boudy, J., and Lockwood, P., "Root Homomorphic Deconvolution Schemes for Speech Processing in Car Noise Environments", Proc. ICASSP'93, pp. 99 - 102, April 1993.

[7] Lockwood, P., Boudy, J., and Blanchet, M., "Non-linear Spectral Subtraction (NSS) and Hidden Markov Models for Robust Speech Recognition in Car Noise Environments", Proc. ICASSP'92, pp. 265 - 268, March 1992.

[8] Lim, J.S., "Spectral Root Homomorphic Deconvolution System", IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 27, pp. 223 - 233, June 1979.

[9] Aurora document no. AU/273/00: "Description and Baseline Results for the Subset of the Speechdat-Car German Database used for ETSI STQ Aurora WI008 Advanced DSR Front-end Evaluation", V1.1, January 12, 2001.

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

[11] Kobayashi, T. and Imai, S., "Spectral Analysis Using Generalized Cepstrum", IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 32, pp. 1087 - 1089, October 1984.