SINGLE- AND TWO-CHANNEL NOISE REDUCTION FOR ROBUST SPEECH RECOGNITION IN CAR

Stefanie Aalburg, Christophe Beaugeant, Sorel Stan, Tim Fingscheidt, Radu Balan, and Justinian Rosca
Siemens AG, ICM Mobile Phones, Grillparzerstrasse 10-18, 81675 Munich, Germany {first_name.last_name@siemens.com}: *Siemens Corporate Research, Multimedia and Video technology, 755 College Road East, Princeton, NJ 08540, USA {Radu.Balan@scr.siemens.com, Justinian.Rosca@scr.siemens.com}*

Abstract:

Hands-free operation of a mobile phone in car raises major challenges for acoustic enhancement algorithms and speech recognition engines. This is due to a degradation of the speech signal caused by reverberation effects and engine noise. In a typical mobile phone/car-kit configuration only the car-kit microphone is used. A legitimate question is whether it is possible to improve the useful signal using the input from the second microphone, namely the microphone of the mobile terminal.

In this paper we show that a speech enhancement algorithm specifically developed for two input channels significantly increases the word recognition rates in comparison with single-channel noise reduction techniques.

Key words: automatic speech recognition, robustness, car noise, two-channel noise reduction

INTRODUCTION

Automatic speech recognition (ASR) systems running in mobile phones can only afford little memory and low computational power, nonetheless they must remain usable under a large variety of noise conditions. Given the fast

growing market for portable devices, robust speech recognition in embedded systems has been enjoying significant attention recently [6].

Besides the quest for robust features [5], [7], [10], two main lines of research aimed at increasing performance of speech recognisers in noise are speech signal enhancement and model adaptation. Although some adaptation techniques [8] achieve very good performance, their use in embedded systems is only of limited interest. This is due to the fact that recognisers operating in mobile phones are subject to constantly changing environments and little or no adaptation data, given that voice control applications consist mostly of short commands which should work on the first try.

In contrast, speech enhancement techniques require no training, therefore they provide "real-time" improvement of recognition rates. For a resource-constrained mobile phone, speech signal enhancement has the added advantage that the same program code can be used to improve not only the recognition rates of the speech recogniser but also the quality of the speech signal for the far-end talker during a voice call. Of course, different tunings of the enhancement algorithm have to be found for both cases in order to optimise for a machine or a human being the listener.

In this paper we compare the performance of single-channel noise reduction algorithms with our newly developed two-channel psycho-acoustically motivated speech enhancement algorithm on an ASR task in the car environment. The novel two-channel speech enhancement algorithm significantly decreases the word error rate for moderate car noise corresponding to city traffic.

SPEECH RECOGNITION SYSTEM

Our recognition engine is based on continuous densities HMM technology with unified variance modelling in Bakis topology, optimised for use in embedded systems to have a small memory footprint and low computational requirements.

For each frame the front-end computes 12 MFCC coefficients plus the total frame energy, as well as their corresponding delta and acceleration values. The inclusion of delta and acceleration coefficients is known to improve robustness of the features against noise. The frames have 32 ms length and an overlap of 17 ms.

The actual feature vector is obtained by retaining the first 24 coefficients of the LDA-transformed super-vector built from two adjacent frames. Finally, the elements of the feature vector are quantised to 8-bit signed values in order to save storage space of the HMM models.

SPECTRAL SUBTRACTION FOR SINGLE-CHANNEL NOISE REDUCTION

Spectral subtraction (SPS) is a frequency-based technique which obtains the clean speech power spectrum $S^2(m, f_k)$ at frame m for the k^{th} spectral component by subtracting the noise spectrum $N^2(m, f_k)$ from the noisy speech $Y^2(m, f_k)$, i.e.

$$S^2(m, f_k) = Y^2(m, f_k) - N^2(m, f_k) \tag{1}$$

Since the noise power spectrum is unknown, an estimated noise power spectrum $\hat{N}^2(m, f_k)$ is subtracted instead, which corresponds to a time varying amplitude filtering of the noisy signal with:

$$H(m, f_k) = 1 - \left(\frac{\hat{N}^2(m, f_k)}{Y^2(m, f_k)} \right)^{1/2} \tag{2}$$

In order to obtain a noise reduction that takes into account the varying noise levels, two factors, the *"over–subtraction factor*" *a* and the *"flooring factor" b*, are introduced:

$$H(m, f_k) := \begin{cases} 1 - a\left(\dfrac{\hat{N}^2(m, f_k)}{Y^2(m, f_k)}\right)^{1/2} & \text{if} \quad 1 - a\left(\dfrac{\hat{N}^2(m, f_k)}{Y^2(m, f_k)}\right)^{1/2} > b \\ b & \text{otherwise} \end{cases}, \qquad (3)$$

The algorithm uses a constant over-subtraction factor $a = 2.5$ and a spectral flooring $b = 0.1$, thus leading to a comparably strong noise suppression for low noise levels.

Since spectral subtraction is the standard solution for noise reduction, it makes a suitable reference for the two algorithms introduced in the following sections.

RECURSIVE LEAST SQUARED AMPLITUDE SINGLE-CHANNEL NOISE REDUCTION

The recursive least square amplitude (RLSA) noise reduction algorithm follows the classical scheme of noise reduction in the frequency domain to estimate the short-time spectral magnitude of the speech $\hat{S}(m, f_k)$ by applying the weighting $G(m, f_k)$ to each short-time Fourier transform coefficient of the noisy speech $Y(m, f_k)$ at frame m:

$$\hat{S}(m, f_k) = G(m, f_k) \cdot Y(m, f_k) \qquad (4)$$

where f_k represents the k^{th} spectral component. Moreover it is assumed that the signal for each frame m can be expressed as the sum of the speech component and the noise component:

$$Y(m, f_k) = S(m, f_k) + N(m, f_k) \qquad (5)$$

The determination of $G(m, f_k)$ involves usually the estimation of the power spectral density (psd) of the noise as well as a rough estimation of the psd of the speech. Most of the time these estimations are empirical and have no link to the definition of $G(m, f_k)$. The recursive implementation of the least-square (LS) criterion has the advantage of deriving the required psd's partially from G [3] and was efficiently used as pre-processing in [1]. It is based on the minimisation of the error function $J_m(e(f_k))$ defined by

$$J_m(e(f_k)) = \sum_{l=0}^{m} \lambda^{m-l} e^2(l, f_k) \tag{6}$$

with λ being a forgetting factor and the errors for each frame l given by:

$$e(l, f_k) = S(l, f_k) - \hat{S}(l, f_k) \tag{7}$$

Considering the vector $\underline{\hat{S}}(m, f_k) = [\lambda^{\frac{m}{2}} \hat{S}(0, f_k); \lambda^{\frac{m-1}{2}} \hat{S}(1, f_k); ...; \hat{S}(m, f_k)]^T$ obtained by the filtering of $\underline{Y}(m, f_k) = [\lambda^{\frac{m}{2}} Y(0, f_k); \lambda^{\frac{m-1}{2}} Y(1, f_k); ...; Y(m, f_k)]^T$ in a similar way as in (4), i.e.

$$\underline{\hat{S}}(m, f_k) = \underline{Y}(m, f_k) \cdot G_{LS}(m, f_k), \tag{8}$$

we obtain the following solution for $G_{LS}(m, f_k)$:

$$G_{LS}(m, f_k) = \frac{\sum_{l=0}^{m} \lambda^{m-l} S^2(l, f_k)}{\sum_{l=0}^{m} \lambda^{m-l} S^2(l, f_k) + \sum_{l=0}^{m} \lambda^{m-l} N^2(l, f_k)} \tag{9}$$

In practical implementation and in order to avoid rough estimation of $S(l, f_k)$ to compute G_{LS}, the following filter is implemented

$$G_{LS}(m, f_k) = \frac{\sum_{l=0}^{m} \lambda_Y{}^{m-l} Y^2(l, f_k)}{\sum_{l=0}^{m} \lambda_Y{}^{m-l} Y^2(l, f_k) + \sum_{l=0}^{m} \lambda_N{}^{m-l} \hat{N}^2(l, f_k)} \tag{10}$$

Where $\hat{N}(l, f_k)$ stands for the noise spectrum estimator and where the different smoothing factor λ_y and λ_N allow us to control the amount of noise reduction and to reduce eventual artefacts on the processed signal.

TWO-CHANNEL NOISE REDUCTION

The two channel noise reduction algorithm [2] is also based on frequency domain analysis. As before, we model our microphone signals $Y_i(m, f_k), \mathrm{i} \in \{1,2\}$ as the sum of speech and noise:

$$Y_i(m, f_k) = S_i(m, f_k) + N_i(m, f_k) \tag{11}$$

Suppose that the speech signal at the microphone 2 can be written as $S_2(m, f_k) = k(f_k) S_1(m, f_k)$, and by introducing the two dimensional vectors $K = [1 \;\; k(f_k)]$, $U = [U_1(m, f_k) \;\; U_2(m, f_k)]$, $U \in \{S, N, Y\}$, Eq. 11 can be rewritten as follows:

$$Y = KS + N \tag{12}$$

An estimation of the speech signal is obtained by a linear filtering of Y through the weighting factor $G_{2mic} = [G_{1,2mic}(m, f_k) \;\; G_{2,2mic}(m, f_k)]$ such that:

$$\hat{S} = G_{2mic} Y^T \tag{13}$$

where Y^T stands for the transpose of the vector Y.

The filter is a two dimensional extension of the algorithm presented in [4], which is effectively a psycho-acoustically motivated weighting rule based on the sound masking properties of the human auditory system. The masking threshold $\gamma_T = [\gamma_{T,1}(m, f_k) \;\; \gamma_{T,2}(m, f_k)]$ is calculated using "clean" speech estimated from spectral subtraction.

The desired amount of noise reduction in the psycho-acoustical sense is defined by a scalar noise attenuation factor $\zeta_n = [\zeta_{n,1} \;\; 0]$. Accordingly, the weighting factor G_{2mic} is chosen in such a way that all components of the residual noise, which exceed the desired amount are just "hidden" below the estimated threshold. It leads to the following solution:

$$G_{2mic} = \zeta_n + \sqrt{\frac{\gamma_T}{K^* \gamma_n^{-1} K}} K^* \gamma_n^{-1} \tag{14}$$

In this last expression γ_n stands for the estimated noise spectral covariance matrix defined by:

$$\gamma_n = \begin{bmatrix} E\left(\left|N_1(m,f_k)\right|^2\right) & E\left(N_1(m,f_k)N_2^*(m,f_k)\right) \\ E\left(N_2(m,f_k)N_1^*(m,f_k)\right) & E\left(\left|N_2(m,f_k)\right|^2\right) \end{bmatrix} \quad (15)$$

with the expectation function $E(.)$. In practice the expectation function is computed as a first order IIR filtering of the estimated noise signal $\hat{N}_1(m,f_k), \hat{N}_2(m,f_k)$.

The vector K is computed by assuming that it is given by an attenuation a and a delay δ :

$$k(f_k) = a.e^{i\delta\omega} \quad (16)$$

Thus the idea is to interpolate the instantaneous ratio

$$\frac{E\left(Y_1(m,f_k)Y_2^*(m,f_k)\right)}{E\left(\left|Y_1(m,f_k)\right|^2\right)}$$

by the model of type Eq. 16. The chosen criteria is the least square criterion in the log domain which give the following formulae:

$$\begin{aligned} \log a &= \frac{2}{N}\sum_{k=0}^{M/2} \log \left| \frac{E\left(Y_1(m,f_k)Y_2^*(m,f_k)\right)}{E\left(\left|Y_1(m,f_k)\right|^2\right)} \right| \\ \delta &= \frac{2}{N}\sum_{k=1}^{M/2} \Im \log \frac{E\left(Y_1(m,f_k)Y_2^*(m,f_k)\right)}{E\left(\left|Y_1(m,f_k)\right|^2\right)} \end{aligned} \quad (17)$$

NOISE ESTIMATION

Noise estimation is required for all three algorithms (spectral subtraction, RLSA, two channel psycho-acoustic). The noise estimators used are based on speech activity detection. The noise spectrum is learned differently during estimated speech/non-speech periods, by smoothing the power spectrum of the current and the previous frame.

Nevertheless, the three algorithms presented here are not based exactly on the same noise estimation rule, even if the principle is similar. Different types of smoothing filter or of voice activity detection criterion are used.

Due to the high interactivity between the weighting rule (Eq. 3, 10 and 14) and the noise estimation rule, the use of a noise estimator unique for the three algorithms is not feasible because each algorithm is tuned in such a way that its overall performance is optimised.

To sum it up, the basic principle of the noise estimation is similar for all three algorithms, but algorithm-specific tuning is necessary in order to obtain optimal results.

DATABASE COLLECTION

Two omni-directional microphones were mounted inside a car, one at the position of a plugged-in mobile phone on the central panel lower-right with respect to the driver (channel 1), and the other one at the lower-left corner of the windshield, facing the driver (channel 2). The car was situated in an acoustic studio and driving noise of varying intensity was played through the two front loudspeakers of the car. This set-up realistically[1] simulates driving noise and reverberation conditions.

Twenty native German speakers, both males and females, were recorded in three different noise conditions:

a) No noise played through the loudspeakers (“clean” speech),
b) With low level noise simulating the driving in city traffic,
c) With high level noise representing highway driving.

Please note that the “clean” speech is actually affected by reverberation effects and other sources of environment noise in the acoustic laboratory.

The utterances consisted of seven isolated German commands and three combination phrases that simulate a dialog between the user and the mobile phone, e.g. “<call> <name> <location>”. For each person and noise level there was a total of ten recordings. The database thus consists of 200 recordings for each noise level, which were used to evaluate the noise reduction algorithms.

[1] Note that we are not simply arithmetically adding the noise to the signal.

EXPERIMENTAL RESULTS

The following table gives the recognition results for the two input channels, once presented in isolation and once in combination.

The first two rows indicate the recognition performance when passing the individual signals of channel 1 and 2 directly to the speech recogniser. The different recognition results are due to the positioning of the microphones. While the car-kit microphone (channel 1) suffers from a degradation of the incoming speech signal due to its position at the centre panel, the windshield microphone (channel 2) remains less affected and outperforms the recognition results obtained with the car-kit microphone. An overall decrease of recognition performance is observed with increased noise level, as expected.

We notice that the position of the microphones strongly influences the word recognition rates for the RLSA and SPS algorithms, which are applied independently to the two channels. Interestingly, channel 1 is the best channel for RLSA, while for SPS channel 2 is better for "clean" speech and low level noise.

Overall RLSA leads to a remarkable improvement of word recognition rates (ranging from +5% to +45% absolute!) outperforming the standard SPS, and is able to level the recognition performance for "clean" speech and speech contaminated with low level car noise.

Passing the combined input of channel 1 and 2 to the recogniser after having applied the two-channel noise reduction, the recognition performance can be improved even further, especially for low level noise. Remarkably, the two-channel noise reduction algorithm performs even better in low level car noise than on "clean" speech, which can be attributed to parameter tuning.

Table 1. [Word Recognition Rates (WRR) for the three noise levels]

Channel number & Noise reduction scheme	"Clean" speech SNR > 20 dB	Low level car noise 14 > SNR >7 dB	High level car noise SNR < 5 dB
Channel 1 (no noise reduction)	80.90 %	55.00 %	30.15 %
Channel 2 (no noise reduction)	86.93 %	68.50 %	35.67 %
Channel 1 noise reduction SPS	88.44 %	73.50 %	63.81 %
Channel 2 noise reduction SPS	90.95 %	75.00 %	51.76 %
Channel 1 noise reduction RLSA	92.45 %	91.00 %	76.88 %
Channel 2 noise reduction RLSA	91.95 %	90.00 %	75.87 %
Two-Channel noise reduction	92.69 %	94.50 %	77.38 %

SUMMARY

In this paper we presented a single-channel noise reduction algorithm (RLSA), based on recursive Wiener filtering in the frequency domain, which leads to an absolute increase of the word recognition rates of up to 45%, and thus outperforms by far the standard spectral subtraction (SPS).

Furthermore, we argued that in a typical mobile phone/car-kit scenario both microphones should be used if better performance of the speech recogniser is desired. To support our claim we presented a newly devised two-channel noise reduction algorithm which performs better than RLSA, yielding a relative decrease of the word error rate of up to 42%.

REFERENCES

[1] B. Andrassy, F. Hilger, C. Beaugeant, *Investigation on the Combination of Four Algorithms to Increase the Noise Robustness of a DSR Front-end for Real World Car Data,* Automatic Speech Recognition and Understanding Workshop, 2001.

[2] R. Balan, J. Rosca and C. Beaugeant, *A Multi-Channel Psycho-acoustically Motivated Noise Reduction Algorithm,* To be published, Sensor Array and Multichannel Signal Processing Workshop, 2002

[3] C. Beaugeant and P. Scalart, *Speech Enhancement Using a Minimum Least Square Amplitude Estimator*, Proc. of 7th Int. Workshop on Acoustic Echo and Noise Control, Darmstadt, Germany, Sep. 10-13, 2001.

[4] S. Gustafsson, P. Jax, and P. Vary, *A Novel Psycho-acoustically Motivated Audio Enhancement Algorithm Preserving Background Noise Characteristics*, Proc. ICASSP, 1998.

[5] L. Jiang and X. Huang, *Acoustic Feature Selection Using Speech Recognisers*, IEEE ASRU Workshop 1999, Keystone, Colorado.

[6] J.-C. Junqua, *Robust Speech Recognition in Embedded Systems and PC Applications*, Kluwer 2000.

[7] K. Kryze, L. Rigazio, T. Applebaum, J-C. Junqua, *A New Noise Robust Sub-band Front-and and its Comparison with PLP*, IEEE ASRU Workshop 1999, Keystone, Colorado.

[8] C. J. Leggetter and P. C. Woodland, *Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models*, Computer Speech and Language, Vol. 9, pp. 171-185.

[9] R. Martin, *Spectral Subtraction Based on Minimum Statistics*, Signal Processing VII: Theories and Applications, EUSIPCO, 1994.

[10] S.D. Peters, P. Stubley, J-M. Valin, *On the Limits of Speech Recognition in Noise*, ICASSP 1999, pp. 365-368.