SOME EXPERIMENTS ON THE USE OF ONE-CHANNEL NOISE REDUCTION TECHNIQUES WITH THE ITALIAN SPEECHDAT CAR DATABASE

M. Matassoni, G.A. Mian†, M. Omologo, A. Santarelli and P. Svaizer

ITC-irst - Centro per la Ricerca Scientifica e Tecnologica
I-38050 Povo - Trento (Italy)
$[matasso, omologo, santarel, svaizer]@itc.it$
†Dipartimento di Elettronica e Informatica - Università degli Studi di Padova (Italy)

ABSTRACT

In this work the use of noise reduction techniques for hands-free speech recognition in car environment is investigated. A set of experiments was carried out using different speech enhancement algorithms based on noise estimation. In particular, linear spectral subtraction and MMSE estimators are considered with various parameter settings.

Experiments were conducted on connected and isolated digits, extracted from the Italian version of the SpeechDat Car database. Recognition rates do not agree with acoustically perceived quality of noise reduction. As a result, the best performance is obtained by spectral subtraction with a suitable choice of the oversubtraction factor and a quantile noise estimator. It provides more than 30% relative performance improvement, from 94.4% of the baseline to 96.2% digit recognition accuracy.

1. INTRODUCTION

Reliable hands-free speech interaction inside the car is still a challenging scenario. An essential requirement is robustness of speech recognition against the various kinds of noise typical of the car environment.

Several new applications in this context are envisaged in the next future, allowing the driver to control by voice devices such as RDS-tuner, CD and cassette player, air conditioner, etc. Also more complex interactions like mobile telephone dialing and access to a navigation system or to remote information services [1] will be practicable in a full hands-free modality, with increased flexibility and safety for the driver who can concentrate his attention on the road.

Security and convenience of hands-free interaction require that the microphone must not encumber the user and therefore can not be put close to his/her mouth. As a consequence the input signal is characterized by a low SNR, being affected by several noise components [2]. Engine and tyres contribute mainly low frequency noise, while aerodynamic turbulence, predominant at high speed, has a broader spectral content [3]. Moreover, other much more unpredictable noise events (e.g. road bumps, rain, traffic noise...) characterize the car environment.

The use of speech enhancement techniques for reduction of environmental noise has been often investigated for speech recognition purposes [4]. Among many others, algorithms based on noise estimation are very popular thanks to their effectiveness and their low computational cost. This paper examines two widely used techniques, namely linear spectral subtraction and MMSE estimators, presenting results of speech recognition on a digit task. One key point of the experiments is the use of the SpeechDat Car database for both training and testing, in order to evaluate system performance under matched conditions. Some experiments were conducted in the past on the use of contaminated speech for system training, as reported in [5]; future activities will be oriented to investigate the use of speech enhancement in a contaminated training framework.

In the following we briefly recall the fundamentals of background noise estimate and the speech enhancement algorithms of interest, then we describe the hands-free speech recognition system under development and summarize the results of our experiments. [1]

2. BACKGROUND NOISE ESTIMATE

The choice of an estimator for background noise is a key problem for noise reduction algorithms. Noise is assumed to be additive and stationary with respect to speech; it is a common practice to estimate it in non-speech intervals and use such estimate to process corrupted speech frames. This approach is based on the use of a VAD (Voice Activity Detector) to separate speech and non-speech intervals, which has proven to be difficult for very noisy conditions.

In recent years, much effort has been spent in search of alternative noise estimators that can work without VAD. In Continuous Spectral Subtraction (CSS, [6]) background noise is estimated at every frame, regardless of speech-non speech classification: this method relies on accurate choice of parameters.

[1] This work was partially funded by the Commission of the EC, Information Society Technologies (IST), 2000-25426, under VICO (Virtual Intelligent COdriver.

In statistical techniques [7, 8, 9], estimation is based on real-time signal statistics, under the assumption that a certain part of the noisy signal power is noise power, also during speech intervals.

2.1. Continuous estimate

The recursive method used in CSS is called here *continuous estimate*:

$$\hat{N}(m,f) = \rho\hat{N}(m-1,f) + (1-\rho)Y(m,f) \qquad (1)$$

where $\hat{N}(m,f)$ is the short-time noise spectrum estimate at m-th frame, and $Y(m,f)$ is the short-time spectrum of the noisy signal at m-th frame and frequency f; parameter ρ is generally chosen close to 1.

Spectra can be magnitude or squared magnitude spectra; in this investigation we used magnitude.

2.2. Quantile estimate

The quantile technique described in [9] was taken as a representative of the statistical estimators class, since in [10] it showed very good performance even when compared to an ideal VAD. A buffer is kept for each frequency component, in order to estimate the q-th quantile of speech distribution. Parameter q was set to 0.5 (*median estimate*) and buffer size was set to 25 according to [9]. To avoid time-consuming sorting operations, we adopted in-place insertion for each new element.

3. NOISE REDUCTION ALGORITHMS

To uniform our notation, X_m will always be the clean speech short-time spectrum, with m indicating the current frame; frequency index is omitted. Further, Y_m will be used for noisy speech spectrum and N_m for noise spectrum. A hat ($\hat{A}$) indicates estimated spectra.

All techniques presented here operate on the short-time amplitude spectrum; noisy spectrum phase is used for reconstruction. Hence, all spectra here are by default amplitude spectra.

All techniques are affected to a various degree by musical noise, the kind of residual distortion which happens with short-time processing [11]; in general, we can assert there is a trade-off between residual noise elimination and speech distortion.

3.1. Spectral subtraction

For each frequency bin, the generalized form of spectral subtraction is defined in [12] as:

$$\hat{X}_m = \max\left(\left[Y_m^{\beta} - \alpha_{SS}\hat{N}_m^{\beta}\right]^{\frac{1}{\beta}}, K_{th}Y_m\right) \qquad (2)$$

where α_{SS} is an over-subtraction factor, β is the power index (β=1 means magnitude subtraction and β=2 means power subtraction), and K_{th} is a noise-floor factor. Over- and under-subtraction can be effective since actual noise is not equal to its mean; on the other hand, noise flooring prevents the occurring of negative values in the spectrum. Actually, in [12] noise floor was a function of the noise power ($K_{th}N_m$ instead of $K_{th}Y_m$), but many authors (e.g. [6]) follow the latter rule, and there is some evidence [13] that it could perform slightly better in a recognition framework. A common value for K_{th} is 0.1.

3.2. MMSE estimators

In [14] and [15] the analytical solution to the problem of the optimal spectrum and log-spectrum MMSE estimator was given, under the hypothesis that spectral components are Gaussian and independently distributed. In both cases the solution is a complex expression of the form

$$\hat{X}_m = G_{opt,m}(\hat{N})Y_m \qquad (3)$$

where it is stressed that the gain function $G_{opt,m}$ depends on present and past values of noise estimate $\hat{N}$; such estimate is used to determine two estimates of *a priori* and *a posteriori* SNR[2]:

$$R_{post,m} = \frac{Y_m^2}{\hat{N}_m^2} \qquad (4)$$

$$R_{prio,m} = (1-\alpha_{EM})R_{post,m} + \alpha_{EM}\frac{\hat{X}_{m-1}^2}{\hat{N}_{m-1}^2} \qquad (5)$$

The value of α_{EM} is generally kept pretty close to 1 for better perceived quality (in [14] the authors suggest to use the value 0.97).

4. EXPERIMENTAL FRAMEWORK

4.1. Acoustic front-end and HMM recognizer

Before feature extraction, the acoustic front-end applies a preemphasis to the input signal. Further, short-time analysis is performed with 20 ms frame size and 16kHz sampling frequency; the analysis step is 10 ms (50% frame overlapping). For each frame, 12 Mel-scaled cepstral coefficients (MCC) and the log-energy are extracted; then the mean value is subtracted from the MCCs, and log-energy is normalized with respect to its maximum value in the utterance. The resulting set of features, together with their first and second order time derivatives, form the 39 components feature vector that is input to the recognizer.

The HMM recognizer is based on a set of 34 phone-like context-independent speech units. Each acoustic-phonetic unit is modeled with left-to-right continuous density HMMs, with output probability distributions represented by means of mixtures having 16 Gaussian components with diagonal covariance matrices. HMM training was accomplished through the standard Baum-Welch training procedure.

All experiments were made under matched conditions: for a given experiment, the training material was processed with the same noise reduction technique applied to the test material.

Note that, with respect to what reported in [5], in this set of experiments a different feature vector was used (12 MCCs instead of 8) and models were not adapted.

[2]Please note that $\hat{N}_m^2$ can be used in place of $\hat{N}_{m-1}^2$ in equation (5) since noise is short-time stationary.

SNR	<5dB	5-15dB	>15dB
number of utterances	403	1137	449

Table 1. Database partitioning according to SNR.

4.2. Digit recognition task

The task considered was extracted from the Italian version of the SpeechDat Car database, which is described in [16]. The task consists of a set of 1084 connected and 905 isolated digits pronounced by 100 speakers, for a grand total of 10771 digits. As shown in Table 1, the test set was divided in three subsets, according to an estimate of SNR.

For a given speech signal, SNR was estimated as $10\log_{10}((P_s - P_n)/P_n)$, where P_s and P_n represent the average power of the segments containing speech and the average power of the background noise segments, respectively; estimation was based on a preliminary manual segmentation of the utterance boundaries.

Training was accomplished using the same material adopted in [5]: it is composed of 2410 phonetically rich utterances pronounced by another subset of 100 speakers.

Performance is measured as digit recognition rate (WRR).

5. EXPERIMENTAL RESULTS

5.1. Baseline

In the following experiments, only one of the three far-microphone channels recorded in SpeechDat Car database collection was used, specifically the rightmost one, placed near the rear mirror (indicated in [5] as *Mic3*). Using the above described front-end, the baseline WRR was 94.4%.

As reference, close-talk recordings (synchronously acquired with far microphones) were used to derive an upper bound performance, equal to 98.9% WRR.

5.2. Spectral subtraction

While in subjective listening spectral subtraction performed poorly, because of a large amount of musical noise, its application to recognition experiments turned out to be beneficial. It was first experimented with the continuous noise estimator, and obtained the best results by setting $\rho = 0.9$ (95.1% WRR). The parameter combination was set to ($\beta = 1, \alpha_{SS} = 1.5, K_{th} = 0.1$), which is frequently used in the literature [3].

After that, the same set of parameters was used with the median noise estimator, obtaining 95.8% WRR. We then chose to explore more deeply the impact of oversubtraction on recognition rates. Our investigation is summarized by Figure 1: the value 1.3 seems to perform slightly better than the others, with a 96.2% recognition rate.

A similar analysis was performed for $\beta = 2$: variations in WRR were distributed on a wider interval, with a maximum around

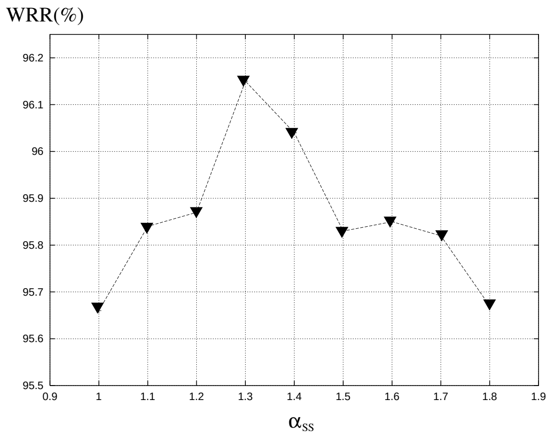

Fig. 1. Recognition rate for spectral subtraction as a function of α_{SS} (with $\beta = 1, K_{th} = 0.1$).

$\alpha_{SS} = 4 \div 4.5$, still under the best performance obtained with magnitude subtraction.

A rough analysis of SNR-dependence for spectral subtraction with median noise estimate was then conducted, by partitioning the results as described in Table 1; the most significant results are shown in Table 2.

	<5dB	5-15dB	>15dB
baseline	90.5	94.4	97.8
$\alpha_{SS} = 1$	92.2	96.1	97.8
$\alpha_{SS} = 1.3$	92.9	96.7	97.7
$\alpha_{SS} = 1.5$	92.2	96.4	97.5
$\alpha_{SS} = 1.7$	92.7	96.2	97.7

Table 2. Performance (WRR%) as a function of SNR.

Table 2 shows that spectral subtraction improves recognition rates for low- and mid-range SNR, while for a good quality input it would be better not to use it at all; the use of a lower α_{SS} (e.g. 1) for high SNR seems to be a good choice as well.

5.3. MMSE estimators

Although MMSE algorithms require heavy computation in their original form, a great amount of computation can be saved if some of the analytical functions involved (e.g. the modified Bessel functions of zero and first order) are approximated by low order polynomials. Thanks to this approach, filtering time was reduced by tenfold (making the algorithm quite as fast as subtraction) while results were almost identical.

However, MMSE spectral amplitude estimator ([14]) does not work well with continuous estimate: the presence of speech components in noise estimate gave rise to “reverberation” tails in speech signal, and recognition rates were far lower than the baseline ones. MMSE log-estimator ([15]), on the other hand, was more pleasant to listen, but its performance was still poor.

The adoption of median estimate gave more significant results, as one can see in Figure 2. The best performance for both techniques appears to be for α_{EM} around 0.88-0.9, at least for the values under investigation.

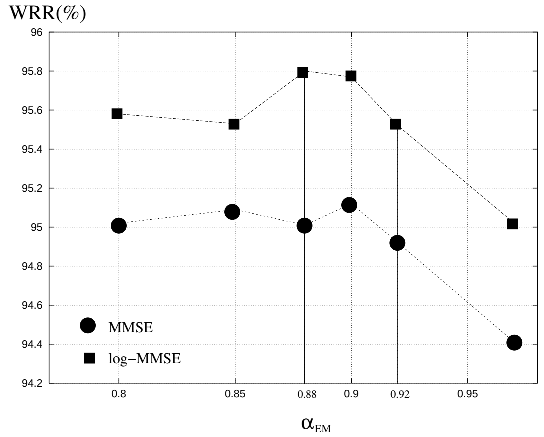

Fig. 2. Comparison of MMSE [14] and log-MMSE [15] estimator performance.

Even though both algorithms give a perceptually more pleasant output than subtraction, recognition rates are lower. Moreover, from our experiments log-spectral estimator appears to be superior to spectral estimator, both for noise reduction and performance improvement.

6. CONCLUSIONS

This work has described an ongoing activity on the adoption of noise reduction techniques in a task of speech recognition in car environment. The study, in its preliminary stage, was conducted using standard algorithms, with different parameter setting, and a real database of digits, which was extracted from SpeechDat Car database. Results showed that spectral subtraction slightly outperforms the other techniques, leading to a relative improvement around 30% in digit accuracy.

Next steps will be oriented to the joint use of the here presented noise reduction techniques and of speech contamination for HMM training [5], possibly with tasks based on medium-large vocabularies. Moreover, the impact of effective speech activity detection algorithms in this framework will be investigated. The final purpose is the development of an advanced system for driver-machine dialogue interaction being studied under VICO project.

7. REFERENCES

[1] Y.Muthusamy, R.Agarwal, Y.Gong and V.Viswanathan, “Speech-enabled information retrieval in the automobile environment”, Proc. ICASSP 1999, vol.4, pp.2259-2262.

[2] M. Omologo, P.Svaizer and M.Matassoni, “Environmental conditions and acoustic transduction in hands-free speech recognition”, Speech Communication, vol.25, pp.75-95, 1998.

[3] P.Lockwood, 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, 1992, pp.215-228.

[4] J-C.Junqua, J-P.Haton, Robustness in automatic speech recognition, Kluwer Academic, 1996.

[5] M.Matassoni, M.Omologo and P.Svaizer, “Use of real and contaminated speech for training of a hands-free in-car speech recognizer”, Proc. of Eurospeech 2001, accepted for publishing.

[6] J.A.Nolazco Flores, S.J.Young, “Continuous speech recognition in noise using spectral subtraction and HMM adaptation”, Proc. ICASSP 1994, vol.I, pp. 409-412.

[7] R.Martin, “Noise power spectral density estimation based on optimal smoothing and minimum statistics”, IEEE Trans. SAP, vol.9, n.5, July 2001, pp.504-512.

[8] H.G.Hirsch, C.Ehrlicher, “Noise estimation techniques for robust speech recognition”, Proc. ICASSP 1995, pp.153-157.

[9] V.Stahl, A.Fischer, R.Bippus, “Quantile based noise estimation for spectral subtraction and Wiener filtering”, Proc. ICASSP 2000, vol.3, pp.1875-1878.

[10] N.W.D.Evans, J.S.Mason, “Noise estimation without explicit speech, non-speech detection: a comparison of mean, median and modal based approaches”, Proc. of Eurospeech 2001, accepted for publishing.

[11] S.V.Vaseghi, Advanced signal processing and digital noise reduction, New York, John Wiley, 1996.

[12] M.Berouti, B.Schwartz and J.Makhoul, “Enhancement of speech corrupted by acoustic noise”, Proc. ICASSP 1979, pp.208-211.

[13] N.W.D.Evans, J.S.Mason, “An assessment of local nonlinear spectral subtraction for remote speech recognition”, Proc. of 1st meeting on Speech Technology, Seville 2000.

[14] Y.Ephraim, D.Malah, “Speech enhancement using optimal non-linear spectral amplitude estimation”, Proc. ICASSP 1983, pp.1118-1121.

[15] Y.Ephraim, D.Malah, “Speech enhancement using a minimum mean-square error log-spectral amplitude estimator”, IEEE Trans. ASSP, vol. ASSP-33, n.2, april 1985, pp.443-445.

[16] L.Cristoforetti, M.Matassoni, M.Omologo, P.Svaizer and E.Zovato, “Annotation of a multichannel noisy speech corpus”, Proc. of LREC, 2000.