IDIAP RESEARCH REPORT

IDIAP

Martigny - Valais - Suisse

A comparison of noise reduction techniques for robust speech recognition

Christopher Kermorvant

IDIAP–RR 99-10

July 1999

Dalle Molle Institute for Perceptual Artificial Intelligence • P.O.Box 592 • Martigny • Valais • Switzerland

phone +41 − 27 − 721 77 11
fax +41 − 27 − 721 77 12
e-mail secretariat@idiap.ch
internet http://www.idiap.ch

IDIAP Research Report 99-10

A comparison of noise reduction techniques for robust speech recognition

Christopher Kermorvant

July 1999

Abstract. This report presents the integration of several noise reduction methods into the front-end for speech recognition developed at IDIAP. The chosen methods are : Spectral Subtraction, Cepstral Mean Subtraction and Blind Equalization. These different methods are studied from a theoretical point of view. Their implementation is described and they are tested on the Numbers95 speech database. A good noise robustness is obtained by combining two of these methods, like Spectral Subtraction with Cepstral Mean Subtraction or Spectral Subtraction with Blind Equalization. The later combination is found to be more appropriate for real recognition systems since it is frame synchronous. A comparison with Jah-RASTA-PLP is also given.

Acknowledgements: The support of the OFES under the grant for the "Speech, Hearing and Recognition" (SPHEAR) project # OFES 970299 is gratefully acknowledged. The work described in this report benefited from fruitful discussions with Chafic Mokbel.

Contents

1 Introduction

Since the early ages of speech recognition, researchers have faced the problem of the degradation of speech recognition system performance when those are used in adverse conditions. This problem is encountered each time a system is used in real-word applications, facing a great diversity of real conditions : background noise, channel interference, microphone distortions, etc. Many solutions have been developed to deal with each problem separately. Classically, these solutions have been classified into two main areas : speech enhancement and models adaptation.

Methods from the first category usually try to remove an estimate of the distortion from the noisy features. For Spectral Subtraction [Bol79], an estimate of the noise spectrum is subtracted from the spectrum of the noisy speech. Similarly, a cepstral bias, like long term cepstral mean, can be computed and subtracted from the cepstral coefficients [Ata74]. In [HMBK92], a high-pass filtering of cepstral coefficients is proposed to remove the effect of the convolutional noise introduced by the channel. Methods based on auditory modeling [Ghi86] have also been proposed in the hope to gain robustness by imitating the naturally robust human perceptual system. All these methods aim at reducing the mismatch between training and testing conditions in the feature space (see Figure 1).

Another approach consists in adapting the recogniser's statistical models to take into account the noise. A combination of two models, one for the speech and another one for the noise was proposed by Varga and Moore [VM90]. In [LLJ91], a Bayesian learning procedure is applied to adapt the parameters of the models. It is worth noting that these approaches have been proven to be superior to speech enhancement approaches [Mok92].

In this report, we describe three speech enhancement methods implemented into the front-end developed at IDIAP for speech recognition. The three methods belong to the first category. They are linear Spectral Subtraction, Cepstral Mean Subtraction and Blind Equalization. They have been chosen for their simplicity and their shown good performance. In the first part of this report, we present the theoretical framework for the three methods. In the second part, we present their performance on the Numbers95 speech database under different noise conditions.

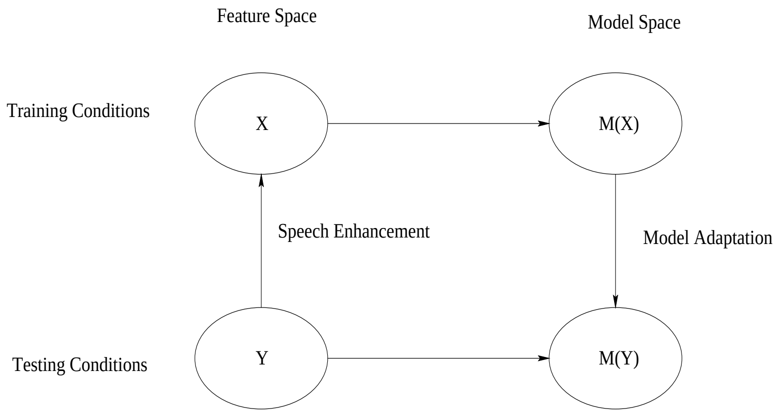

Figure 1: *Mismatch in training and testing conditions*

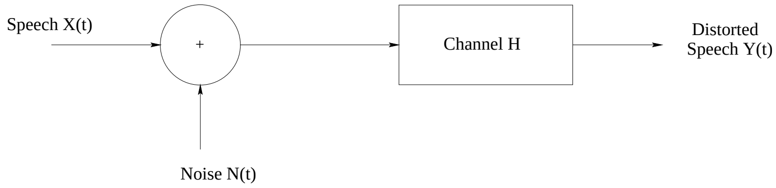

Figure 2: *Schematic view of telephone network distortions*

2 Noise Reduction Techniques

In this section, we describe the noise reduction techniques we have tested : Spectral Subtraction, Cepstral Mean Subtraction and Blind Equalization. Both the theory underlying each method and the practical aspects of the implementation are given.

2.1 Spectral Subtraction

2.1.1 Method

Let us consider a speech signal $s(t)$ degraded by uncorrelated additive noise $n(t)$. The resulting signal is then

$$y(t) = s(t) + n(t) \tag{1}$$

If we consider $Y(\omega), S(\omega)$ and $N(\omega)$ the respective Fourier transforms of $y(t), s(t)$ and $n(t)$, we obtain

$$\mid Y(\omega) \mid^2 = \mid S(\omega) \mid^2 + \mid N(\omega) \mid^2 + S(\omega) \cdot N^*(\omega) + N(\omega) \cdot S^*(\omega) \tag{2}$$

where $N^*(\omega)$ and $S^*(\omega)$ are the complex conjugates of $N(\omega)$ and $S(\omega)$.

We want to obtain an estimate of $\mid S(\omega) \mid^2$ which is the short-time energy of clean speech. In Equation 2, only $\mid Y(\omega) \mid^2$ can be directly derived from the observed data. However $\mid N(\omega) \mid^2, S(\omega) \cdot N^*(\omega)$ and $N(\omega) \cdot S^*(\omega)$ can be approximated respectively by $E[\mid N(\omega) \mid^2], E[S(\omega) \cdot N^*(\omega)]$ and $E[N(\omega) \cdot S^*(\omega)]$, where $E[.]$ denote the ensemble average. With the hypothesis that the noise $n(t)$ is uncorrelated with speech $s(t)$, we have $E[S(\omega) \cdot N^*(\omega)]$ and $E[N(\omega) \cdot S^*(\omega)]$ equal to zero and we can thus derive an estimate $\mid \widehat{S}(\omega) \mid^2$ of $\mid S(\omega) \mid^2$:

$$\mid \widehat{S}(\omega) \mid^2 = \mid Y(\omega) \mid^2 - E[\mid N(\omega) \mid^2] \tag{3}$$

$E[\mid N(\omega) \mid^2]$ is estimated on signal, during non-speech periods. Note that Eq. 3 corresponds to the Power Spectral Subtraction and is equivalent to a Wiener filter.

It is important to note that Eq. 3 does not guarantee that $\mid \widehat{S}(\omega) \mid^2$ is positive. Negative value of $E[\mid N(\omega) \mid^2]$ should be set to zero or better to a constant non-zero minimum value. The main drawback of this method is that it introduces non-linearities in the spectrum , known as musical noise, which are very harmful for speech recognition.

A generalization of the formula 3 was proposed by Berouti et *al.* [BSM79]. In this approach, the spectrum is raised to a power a. Then if we define $T(\omega)$ as

$$T(\omega) = \mid Y(\omega) \mid^{2a} - \alpha(SNR) * \mid \widehat{N}(\omega) \mid^{2a} \tag{4}$$

where $\alpha(SNR)$ is an SNR-dependent noise overestimation factor, the estimate of clean speech is given by :

$$\mid \widehat{S}(\omega) \mid^2 = \begin{cases} T(\omega)^{1/a} & \text{if } T(\omega)^{1/a} > \beta \mid N(\omega) \mid^2 \\ \beta \mid N(\omega) \mid^2 & \text{otherwise} \end{cases} \tag{5}$$

The overestimation factor is supposed to reduce the musical noise introduced by the subtraction and β defines the minimum spectral value after subtraction.

Another enhancement of this technique has been developed by Lockwood and Boudy [LB91][LB92]. The basic idea of this approach is to subtract a minimum of noise at high SNR and to remove a maximum of noise at low SNR. The subtracted factor not only depends on the estimated noise $E[\mid N(\omega) \mid^2]$ but also on the estimated SNR and on a frequency dependent factor $\alpha(\omega)$. Eq. 3 becomes :

$$\mid \widehat{S}(\omega) \mid^2 = \mid Y(\omega) \mid^2 - \Phi(SNR, \alpha(\omega), E[\mid N(\omega) \mid^2]) \tag{6}$$

Several functions Φ are proposed in [LB92], for example :

$$\Phi(\omega) = \alpha(\omega) - sigmoid(SNR(\omega))\Big(\alpha(\omega) - E[\mid N(\omega) \mid^2]\Big) \tag{7}$$

However, since this function needs a lot of parameter tuning, we decided not to develop this approach.

2.1.2 Implementation

The spectral subtraction needs an estimate of the noise power spectrum. This estimate is generally computed during non-speech periods. To determine non-speech periods, a statistical distance is computed for each frame between the spectrum of the signal and the distribution of the noise [MMK+97]. This distance is compared with a threshold to decide whether or not this frame corresponds to a non-speech period. If the spectrum corresponds to a non-speech period, the noise characteristics (mean and variance) are updated with a first order adaptive process, with factors α and $1-\alpha$. A typical value for α is 0.99, which corresponds to an adaptation over 100 frames, that is 1 second. The initialization of the noise estimate is done on the first 10 frames (this makes the assumption that the first 10 frames contain only noise).

The problem of non-linearities introduced by the flooring in the spectral subtraction is one of the major problem of this method. To deal with this problem, Boll [Bol79] proposed to replace $\mid Y(\omega) \mid^2$ in formula 3 by an average over 3 frames, $\overline{\mid Y(\omega) \mid^2}$. We chose to smooth the subtracted spectrum $\mid \widehat{S}(\omega) \mid^2$ with a low-pass filtering in order to remove the non-linearities induced by the flooring. We obtain a smoothed estimate of the clean speech spectrum $\mid \bar{S}_t(\omega) \mid^2$ with :

$$\mid \bar{S}_t(\omega) \mid^2 = \gamma * \mid \bar{S}_{t-1}(\omega) \mid^2 + (1-\gamma) * \mid \widehat{S}_t(\omega) \mid^2 \tag{8}$$

Several values have been tested for γ and the best recognition results have been obtained with $\gamma = 0.5$.

2.2 Cepstral Mean Subtraction

2.2.1 Method

Cepstral Mean Subtraction is one of the earliest and the simplest methods used to remove channel distortion from signal. The principle behind this method is that a convolutional distortion in the time domain, such as a channel distortion, corresponds to an additive distortion in the cepstral domain. If we denote by $s(t)$ a speech signal, by $w(t)$ the channel impulse response and by $y(t)$ the speech signal transmitted through the channel we have the following equivalence :

$$y(t) = s(t) \otimes w(t) \Leftrightarrow C_y(i) = C_s(i) + C_w(i) \tag{9}$$

where $\otimes$ is the convolution operator and $C_y(i)$, $C_s(i)$ and $C_w(i)$ are the cepstrum of respectively the transmitted signal, the speech signal and the channel . Now if we apply the expectation operator to the right side of the equivalence, we have :

$$\overline{C_y(i)} = \overline{C_s(i)} + \overline{C_w(i)} \tag{10}$$

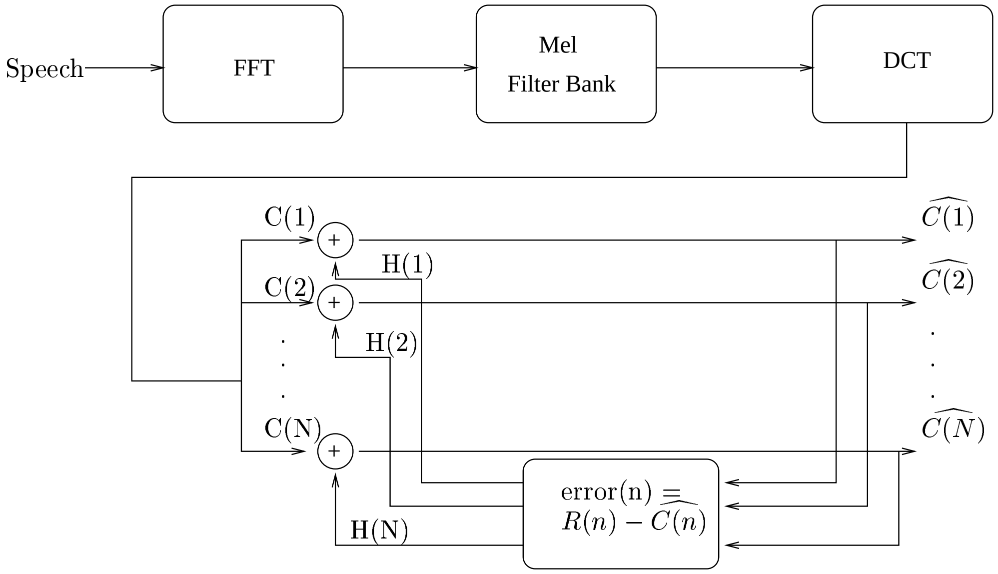

Figure 3: Blind equalization in the cepstral domain (after Mauuary [Mau98])

With the hypothesis that the channel characteristics are constant and that the expectation of the speech cepstrum is null (except for the 0th coefficient), we obtain

$$\overline{C_y(i)} = C_w \tag{11}$$

Now by computing the long time average of the cepstrum of the transmitted speech, we have :

$$C_w = \frac{1}{N} \sum_{i=1}^{N} C_y(i) \tag{12}$$

It is now possible to subtract C_w from the observed cepstral vectors $C_y(i)$ in order to remove the channel effect.

2.2.2 Implementation

The Cepstral Mean Subtraction is a noise reduction technique very simple to implement. The long-time average of the cepstral coefficients is computed off-line and subtracted from each coefficient. This subtraction is done for all coefficients except the coefficient corresponding to the energy.

2.3 Blind Equalization

2.3.1 Method

The blind equalization is a filtering technique which has been studied in digital communication in order to remove the channel effect through the sole observation of the output of the channel. An error criterion based on known statistics of the transmitted signal is used to compute the parameters of the filter. This can be implemented with an adaptive filter [Shy92], as it was proposed by Mokbel et *al.* [MJM96]. This technique can be applied either in the spectral domain [Mau96] or in the cepstral domain [Mau98].

If we consider a speech signal $x(t)$ transmitted over a channel with impulse response $w(t)$, the resulting signal $y(t)$ is given by :

$$y(t) = x(t) \otimes w(t) \tag{13}$$

where $\otimes$ is the convolution operator. The basic idea is to apply a filter $h(t)$ to the observed signal $y(t)$ in order to obtain an estimate of the original speech signal, $\hat{x}(t)$:

$$\hat{x}(t) = y(t) \otimes h(t) \tag{14}$$

Eq. 13 can be written in the frequency domain:

$$\Gamma_y(f) = \Gamma_x(f)W^2(f) \tag{15}$$

so that Eq. 14 becomes :

$$\Gamma_{\hat{x}}(f) = \Gamma_x(f)W^2(f)H^2(f) \tag{16}$$

where $\Gamma_{\hat{x}}(f)$ and $\Gamma_x(f)$ are the spectral densities of $\hat{x}(t)$ and $x(t)$ respectively , and where $W(f)$ and $H(f)$ are the transfer function of the channel and of the adaptive filter impulse response respectively. Similarly, Eq. 14 can be derived in the cepstral domain to obtain:

$$C_{\hat{x}}(i) = C_x(i) + C_W(i) + C_H(i) \tag{17}$$

where $C_{\hat{x}}(i)$ and $C_x(i)$ are the cepstrum of the equalized speech and the original speech respectively and where $C_W(i)$ and $C_H(i)$ are the cepstrum of the channel and of the adaptive filter respectively.

The parameters of the filter are adapted using some knowledge about the signal. This knowledge can be either the value of the long-term average of speech spectrum for an adaptation in the spectral domain or the value of the long-term average of speech cepstrum for an adaptation in the cepstral domain. In the following, we present the adaptation of the filter parameters only in the cepstral domain. This adaptation is similar in the spectral domain.

The error between the long-term average of speech cepstrum $R(i)$ and the equalized speech $C_{\hat{x}}(i)$ is

$$error(i) = R(i) - (C_x(i) + C_W(i) + C_H(i)) \tag{18}$$

We can use an minimum mean square error criterion (MMSE) to determine the optimal value for C_H. The mean square error is :

$$E[error^2(i)] = E[(R(i) - (C_x(i) + C_W(i) + C_H(i))^2] \tag{19}$$

which can be minimized to obtain the optimal filter:

$$E[C_H(i)] = R(i) - E[C_x(i)] - C_W(i) \tag{20}$$

If we consider that $R(i) - E[C_x(i)]$ is equal to zero, which means that the long-time average of the initial speech is really equal to the expected long-time average, we obtain the optimal values for the filter parameters which remove the channel effect.

2.3.2 Implementation

The filter parameter values can be computed with a simple adaptive formula [Mau98] :

$$C_H^{n+1}(i) = C_H^n(i) + \mu(R(i) - (C_H^n(i) + C_y^n(i)) \tag{21}$$

where μ is the adaptation coefficient and the upper script n denotes the frame number. The convergence of $C_H^n(i)$ is strongly dependent on μ : μ must be large enough to ensure a quick convergence but small enough to avoid the disturbance of short-time variations of speech cepstrum values. We used $\mu = 0.005$, which was the best value found in [Mau98]. Some improvements of the convergence can be obtained by giving more importance to high energy frames compared to low energy frames [MJM96]. For this purpose, we used a piecewise linear function to adapt the value of μ according to the frame energy $E(n)$:

$$\mu(E(n)) = \begin{cases} 0.5 & \text{if } 0 < E(n) \leq 350 \\ 0.2 * (E(n) - 350) + 0.5 & \text{if } 350 < E(n) \leq 350 \\ 2.5 & \text{if } 450 < E(n) \end{cases} \tag{22}$$

The parameters of this piecewise linear function where computed on the training database.

3 Evaluation of the performance

3.1 Database

The speech database chosen for the experiments is the Numbers'95 database [CNLD95] from the Center for Spoken Language Understanding (CSLU). This database contains digits sequences continuously spoken over the telephone. The database consists of 10.5 hours of speech and has been separated into 3 sets : the training set, the development-test set and the test set. We used the 3590 sentences of the training set for the training of our models and we tested them on the 1206 sentences of the development-test set.

3.2 Feature Extraction

We used the front-end developed at IDIAP to extract the feature vectors from the speech files. We computed 26 mel-scaled filter bank coefficients, over a 32 ms hamming window, with a 10 ms shift. Then 13 mel-cepstral coefficients were derived together with their first and second order derivatives (for a total of 39 coefficients).

3.3 Models

The recognition system was based on Gaussian mixture HMMs. It was trained with HTK [You97]. The system was composed of 81 triphones modeled by 3 states HMMs; each state had a 10 Gaussian mixture pdf and a diagonal covariance matrix. No language model was used. The training sequence started with 1 Gaussian monophones initialized with the segmented training set (HINIT), followed by gaussian splitting and embedded models re-estimation (HEREST) up to 10 Gaussian triphones. At this point, the Word Entrance Penalty (WEP) parameter was tuned. Several recognitions experiments were done on the standard development-test set with different WEP values and the value leading to the best recognition score was kept.

Note that we trained a different system for each speech enhancement methods that we tested.

3.4 Experiments Setup

For the testing, the task was to recognize the 1206 sentences from the Numbers95 database development-test under different simulated noise conditions. We used three kinds of recorded noise from the NOISEX92 database [VSTJ92]: car noise, factory noise,and lynx helicopter noise. The noise was added to the clean speech at different SNR levels (18 db, 12 db, 6db, 0db). In the following results, the clean speech is presented at 30 dB. The SNR was computed excluding silence, at utterance level.

Since the Spectral Subtraction relies on an estimate of the noise spectrum, it is possible to evaluate the performance of the subtraction scheme independently from the noise estimate by using the *a priori* spectrum of the noise. This spectrum is computed directly on the noise signal, before it is added to the speech signal. The recognition results using the *a priori* values of the noise spectrum are also presented.

The recognition results are given in percent of word error rate (WER). The confidence interval was computed for a given WER p with the standard formula :

$$i = d_\Omega * \sqrt{\frac{p * (100 - p)}{N}} \qquad (23)$$

with in our case $d_\Omega = 1.96$ (for a confidence interval at 95%) and $N = 4670 * 13$ (the number of words in the development-test set times the number of noise conditions).

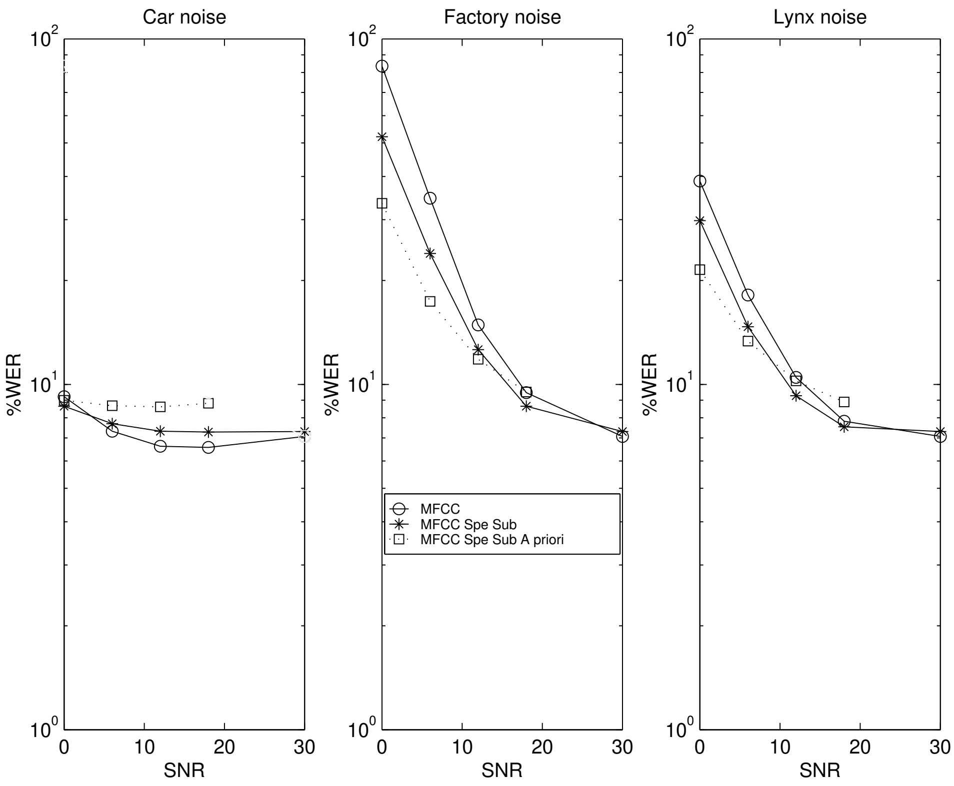

Figure 4: Results with Spectral Subtraction

3.5 Recognition Results

3.5.1 Spectral Subtraction

The evaluation of the Spectral Subtraction performance under the three noise conditions is shown in Figure 4 and is summarized in the following table (see also Appendix A):

	Mean WER	Confidence interval at 95%
MFCC	19.57	19.25 - 19.89
MFCC Spectral Subtraction	15.14	14.85 - 15.42
MFCC Spectral Subtraction A Priori	13.43	13.16 - 13.70

The first point to note is that using the Spectral Subtraction can decrease the performance on clean or slightly noisy speech . Even if the difference is not significant (7.30% WER with Spectral Subtraction compared to 7.07% without with a confidence interval of 0.74%), it is worth noting that even in some noisy conditions using the Spectral Subtraction significantly decreases the performance. For example until 6 dB of SNR for car noise, using Spectral Subtraction increases the WER. This is due to the non-linearities induced by the Spectral Subtraction technique. This effect occurs since the Numbers95 database is a real telephone speech database : even if there is no noise added to the speech, the noise estimated on the first milliseconds of signal is not null, due to the recording conditions. Second, the positive effect of the noise subtraction can be seen as from 18 dB of SNR for factory and lynx noise. On average, for the three noise condition, using Spectral Subtraction yields a relative decrease of error rate of 22.6%. Finally, the importance of the noise estimate can be seen by

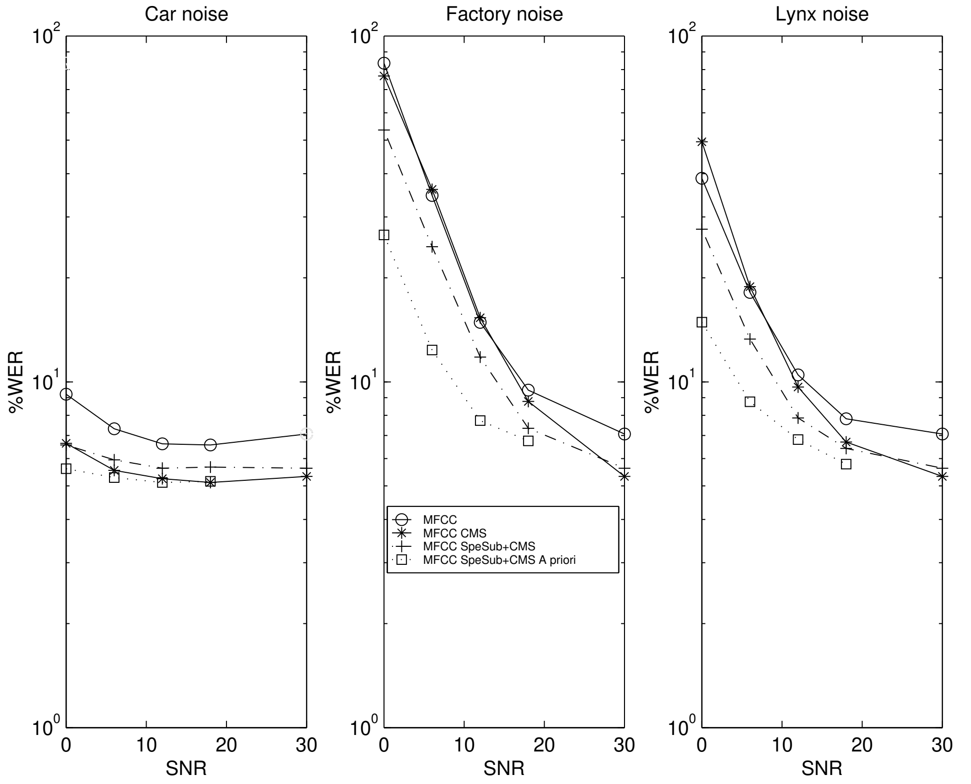

Figure 5: Results with Cepstral Mean Subtraction

comparing the results between the Spectral Subtraction and the *a priori* Spectral Subtraction. When the noise level is high, having a good estimate of the noise spectrum is more important : with the lynx noise at 0 dB of SNR the error rate is reduced from 29.79% to 21.48% with the *a priori* noise spectrum. This effect is even more important when the noise is unstationnary, like factory noise : at 0 dB of SNR, the error rate is reduced from 52.10% to 33.43%.

3.5.2 Cepstral Mean Subtraction

The evaluation of the Cepstral Mean Subtraction performance under the three noise conditions is shown in Figure 5 and is summarized in the following table (see also Appendix A):

	Mean WER	Confidence interval at 95%
MFCC	19.57	19.25 - 19.89
MFCC CMS	19.17	18.87 - 19.49
MFCC Spectral Subtraction + CMS	13.99	13.72 - 14.27
MFCC A Priori Spectral Subtraction + CMS	9.24	9.01 - 9.47

First, the Cepstral Mean Subtraction significantly reduce the error rate on clean speech from 7.07% to 5.33 % (24.6% reduction). This result was expected since Numbers95 is a telephone speech database and since the Cepstral Mean Subtraction removes the convolutional channel noise. Second, the combination of both Spectral Subtraction and Cepstral Mean Subtraction also gives significant improvement in noise conditions. These two methods combined decrease the error rate for the three noise condition and at all SNR level. On average, for the three noise conditions, using both Spectral

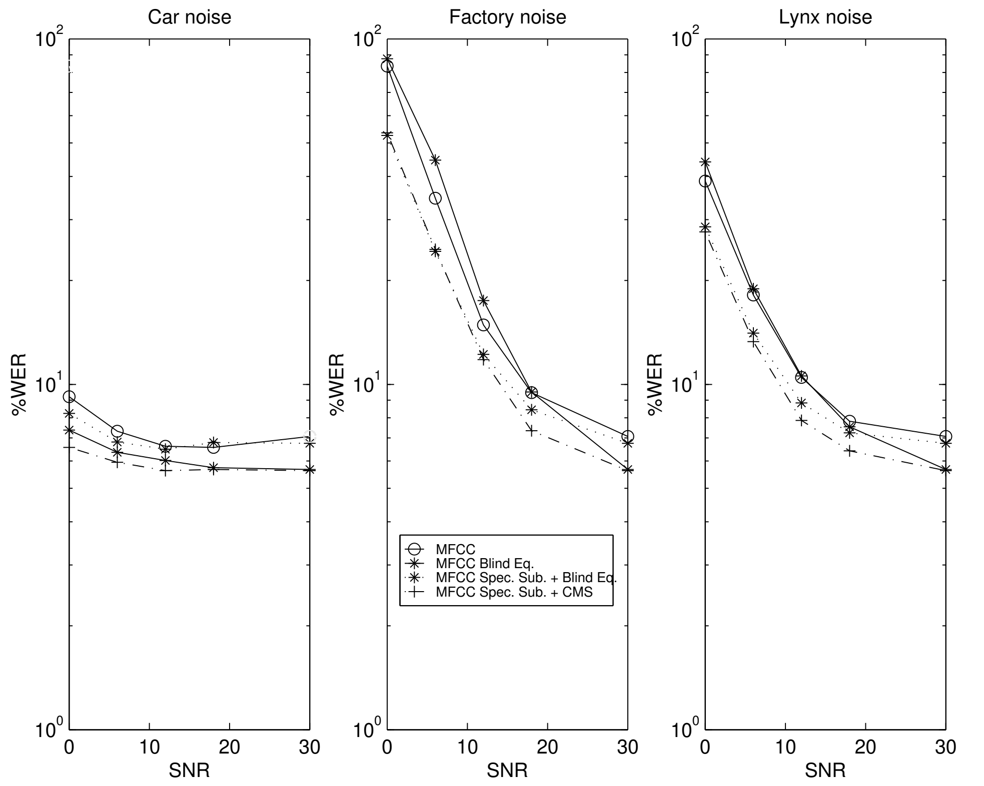

Figure 6: Results with Blind Equalization

Subtraction and Cepstral Mean Subtraction yields a relative decrease of error rate of 28.5 % (from 19.57% to 13.99%). Finally, the importance of the noise estimate can be evaluated by using the *a priori* noise spectrum. Using both Spectral Subtraction and Cepstral Mean Subtraction with the *a priori* known noise spectrum reduces the error rate on average for the three noises from 52.8%. This very good result reveals the potential of the combination of these two methods. Indeed, we could improve the recognition results obtained with the current techniques with better noise estimates.

3.5.3 Blind Equalization

The evaluation of the Blind Equalization performance under the three noise conditions is shown in Figure 6 and summarized in the following table (see also Appendix A) :

	Mean WER	Confidence interval at 95%
MFCC	19.57	19.25 - 19.89
MFCC Blind Eq.	20.88	21.21 - 20.56
MFCC Spectral Subtraction + Blind Eq.	14.71	14.43 - 14.99
J-RASTA-PLP	15.29	15.00 - 15.57

The first conclusion is that the Blind Equalization reduces the recognition error rate on clean speech. From 7.07% error rate on clean speech with the baseline system, using Blind Equalization lowers the error rate to 5.67% which represents 20% of error rate reduction. This reduction is due to the fact that the Numbers95 database is a telephone speech database, and that the Blind Equalization removes the convolutional channel noise. Secondly, when combined with the Spectral Subtraction, the

Blind Equalization provides a good robustness to additive noise. On average for all noise conditions, using both Spectral Subtraction and Blind Equalization reduces the recognition error rate of 27.4%. This reduction should be compared to the reduction obtained with Cepstral Mean Subtraction. Even if Cepstral Mean Subtraction seems to yield more error rate reduction (28.5%), the main advantage of the Blind Equalization is that it is frame synchronous, and then suitable for real systems.

Blind equalization has been compared with J-RASTA-PLP features. Indeed, J-RASTA-PLP are robust features which are designed to remove both additive noise and convolutional noise [HM94]. 12 J-RASTA coefficients were computed with the following parameters : a window length equals to 25 ms, a window shift equals to 12.5 ms, the LPC analysis order equals to 10. The results in the previous table show that using Blind Equalization with Spectral Subtraction yields better results than using J-RASTA-PLP. On average for the three noises, using Blind Equalization with Spectral Subtraction reduces the error rate from 27.4% whereas using J-RASTA-PLP only reduces the error rate from 21.9%.

4 Conclusion

In this report, we described several noise reduction techniques implemented and tested at IDIAP : Spectral Subtraction, Cepstral Mean Subtraction and Blind Equalization. Both additive and convolutional noise can be removed by these techniques : Spectral Subtraction removes additive noise and Cepstral Mean Subtraction and Blind Equalization removes convolutional noise. On the Numbers95 database under three noise conditions, we obtained a robustness to noise by combining two of these methods like Spectral Subtraction and Cepstral mean Subtraction or Spectral Subtraction and Blind Equalization. However, the combination of Spectral Subtraction and Blind Equalization has the main advantage of being frame synchronous. The experiments also showed that the methods we developed are better than another equivalent noise reduction techniques, namely J-RASTA-PLP.

References

[Ata74] B. S. Atal. Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification. *JASA*, 55(6):1304–1312, June 1974.

[Bol79] Steven F. Boll. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Trans. on Acoustics, Speech and Signal Processing*, 27(2), April 1979.

[BSM79] N. Berouti, R. Schwartz, and J. Makhoul. Enhancement of speech corrupted by acoustic noise. In *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, pages 208–211, 1979.

[CNLD95] R.A. Cole, M. Noel, T. Lander, and T. Durham. New telephone speech corpora at cslu. *Proc. European Conf. on Speech Communication and Technology*, 1:821–824, 1995.

[Ghi86] O. Ghitza. Auditory nerve representation as a front-end for speech recognition in a noisy environment. *Computer Speech and Language*, 2:109–130, 1986.

[HM94] H. Hermansky and N. Morgan. RASTA processing of speech. *TransASSP*, 2(4):578–589, October 1994.

[HMBK92] H. Hermansky, N. Morgan, A. Bayya, and P. Kohn. Rasta–plp speech analysis technique. *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, 1:121–124, 1992.

[LB91] P. Lockwood and J. Boudy. Experiments with a non-linear spectral subtractor (nss), hidden markov models and the projection, for robust speech recognition in car. In *Proc. European Conf. on Speech Communication and Technology*, pages 79–82, 1991.

[LB92] P Lockwood and J. Boudy. Experiments with a nonlinear spectral subtractor (nss), hidden markov models and the projection, for robust speech recognition in cars. *Speech Communication*, 11:215–228, 1992.

[LLJ91] C.H. Lee, C. H. Lin, and B.H. Juang. A study on speaker adaptation of the parameters of continuous density hidden markov models. *IEEE Trans. on Signal Processing*, 39(4), 1991.

[Mau96] L. Mauuary. Blind equalization for robust telephone based speech recognition. In *Proc. European Signal Processing Conference*, 1996.

[Mau98] L. Mauuary. Blind equalization in the cepstral domain for robust telephone based speech recognition. In *Proc. European Signal Processing Conference*, 1998.

[MJM96] C. Mokbel, D. Jouvet, and J. Monn. Deconvolution of telephone line effects for speech recognition. *Speech Communication*, 19:185–196, 1996.

[MMK+97] C. Mokbel, L. Mauuary, L. Karray, D. Jouvet, J Monne, J. Simonin, and K. Bartkova. Towards improving asr robustness for psn and gsm telephone applications. *Speech Communication*, 23:141–159, 1997.

[Mok92] C. Mokbel. *Reconnaissance de la Parole dans le Bruit : Bruitage/Débruitage*. PhD thesis, Ecole Nationale Supérieure des Télécommunications, 1992.

[Shy92] J. J. Shynk. Frequency-domain and multirate adaptive filtering. *IEEE Signal processing magazine*, pages 15–37, January 1992.

[VM90] A. P. Varga and R. K. Moore. Hidden Markov model decomposition of speech and noise. In *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, pages 845–848, 1990.

[VSTJ92] A. Varga, H.J.M. Steeneken, M. Tomlinson, and D. Jones. The noisex–92 study on the effect of additive noise on automatic speech recognition. *Technical Report, DRA Speech Research Unit*, 1992.

[You97] Steve Young. *The HTK Book*. Cambridge University, March 1997.

A Appendix : Summary of recognition results

A.1 Recognition error rate on Numbers95 with added car noise for different noise reduction techniques

	SNR 0	SNR 6	SNR 12	SNR 18	Clean
MFCC	9.21	7.32	6.62	6.57	7.07
MFCC SpeSub	8.65	7.69	7.32	7.28	7.30
MFCC SpeSub Apriori	8.99	8.67	8.61	8.82	-
MFCC CMS	6.64	5.55	5.25	5.12	5.33
MFCC SpeSub CMS	6.57	5.95	5.63	5.67	5.63
MFCC SpeSub CMS apriori	5.61	5.29	5.12	5.16	-
MFCC Blind Eq.	7.37	6.36	6.02	5.74	5.67
MFCC SpeSub Blind Eq.	8.24	6.81	6.49	6.79	6.75
MFCC SpeSub Blind Eq. apriori	6.75	6.36	6.60	6.60	-
J-RASTA	7.71	7.28	7.47	7.52	7.43

A.2 Recognition error rate on Numbers95 with added factory noise for different noise reduction techniques

	SNR 0	SNR 6	SNR 12	SNR 18	Clean
MFCC	83.47	34.60	14.86	9.46	7.07
MFCC SpeSub	52.10	23.92	12.61	8.65	7.30
MFCC SpeSub Apriori	33.43	17.39	11.82	9.51	-
MFCC CMS	76.68	36.00	15.33	8.78	5.33
MFCC SpeSub CMS	53.51	24.58	11.80	7.34	5.63
MFCC SpeSub CMS apriori	26.60	12.36	7.73	6.75	-
MFCC Blind Eq.	87.64	44.60	17.49	9.51	5.67
MFCC SpeSub Blind Eq.	52.55	24.26	12.21	8.44	6.75
MFCC SpeSub Blind Eq. apriori	28.52	13.75	8.69	7.17	-
J-RASTA	50.43	24.05	12.85	8.93	7.43

A.3 Recognition error rate on Numbers95 with added lynx noise for different noise reduction techniques

	SNR 0	SNR 6	SNR 12	SNR 18	Clean
MFCC	38.80	18.14	10.47	7.82	7.07
MFCC SpeSub	29.79	14.69	9.27	7.54	7.30
MFCC SpeSub Apriori	21.48	13.34	10.24	8.89	-
MFCC CMS	49.44	18.84	9.66	6.70	5.33
MFCC SpeSub CMS	27.64	13.30	7.86	6.42	5.63
MFCC SpeSub CMS apriori	14.88	8.76	6.81	5.78	-
MFCC Blind Eq.	44.05	18.92	10.58	7.54	5.67
MFCC SpeSub Blind Eq.	28.57	14.07	8.84	7.24	6.75
MFCC SpeSub Blind Eq. apriori	14.50	9.40	7.56	6.66	-
J-RASTA	31.35	15.95	9.85	7.97	7.43