A COMPARISON OF TWO STRATEGIES FOR ASR IN ADDITIVE NOISE : MISSING DATA AND SPECTRAL SUBTRACTION

Christopher Kermorvant and Andrew Morris

IDIAP
P.O.Box 592, Martigny, Switzerland.
kermorva,morris@idiap.ch

ABSTRACT

This paper addresses the problem of speech recognition in the presence of additive noise. To deal with this problem, it is possible to estimate the noise characteristics using methods which have previously been developed for speech enhancement techniques. Spectral subtraction can then be used to reduce the effect of additive noise on speech in the spectral domain. Some techniques have also recently been proposed for recognition with missing data. These approaches require an estimation of the local SNR to detect the speech spectral features which are relatively free from noise so as to perform recognition on these parts only. In this article, we compare these two different strategies, spectral subtraction and ”missing data”, on continuous speech additively disturbed with real noise. It is shown that missing data methods can improve recognition performance under certain noise conditions but still need to be improved in order to to reach the performance of the spectral subtraction.

1. INTRODUCTION

The problem of reducing the performance degradation of speech recognition systems in the presence of additive noise has been investigated for several years. Overall, researchers have tried to make recognition systems noise resistant in three main ways: first by using noise resistant features (e.g. systems based on spectral subtraction [1]); second, by adapting the recogniser's statistical models to noise (e.g. by using parallel model combination [6]; third, by using a distance measure that is robust to noise (e.g. [4]). Recently, some techniques have been proposed based on missing data theory [5]. These techniques try to include knowledge about the noise level in the way likelihoods are computed.

In this paper, we compare this missing data approach with spectral subtraction. Both of these methods need an estimate of the noise level, as described in Section 2. The missing data approach and the spectral subtraction are described in Section 3 and 4 respectively. Tests are described and discussed in Sections 5,6,and 7. Finally, conclusions and possible improvements are given in Section 8.

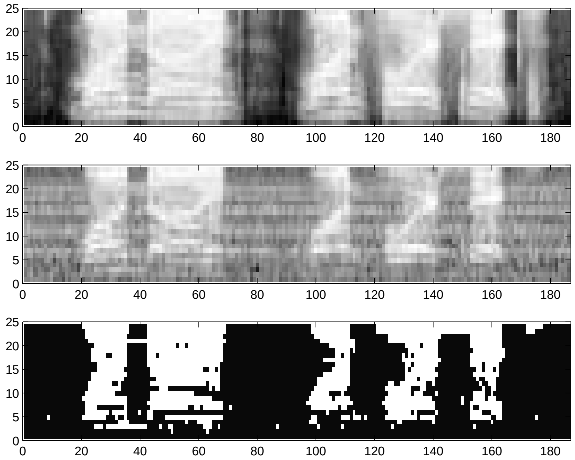

Figure 1: *Clean filter bank coefficients (top), noisy filter bank coefficients (with lynx noise at 6 dB) (middle) and mask (bottom)*

2. NOISE ESTIMATION

Both of the techniques that we compare in this paper need an estimate of the noise spectrum. Typically, this estimate is computed on the signal, during the non-speech periods. If we consider a speech signal $s(t)$ that is degraded by additive noise $n(t)$, the resulting signal is then

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

During non-speech periods, $s(t)$ equals zero so that the spectrum of $y(t)$, $\mid Y(\omega) \mid^2$, is composed only of the noise. We can use these values to compute the spectral characteristics of the noise, $\mid \widehat{N}(\omega) \mid^2$, modeled by a gaussian density, defined by its mean and variance.

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. 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 factor (α and $1 - \alpha$). The initialisation of the noise estimate is done on the first 10 frames (this makes the assumption that the first 10 frames contain only noise).

3. MISSING DATA RECOGNITION

In the case of speech recognition in additive noise conditions, we can consider that some components of the feature vectors in the spectral domain are masked by noise and can thus be seen as missing. The spectral feature vectors can be separated into two parts: the components which are present (reliable) and those which are missing (unreliable) : $x = (x^p, x^m)$. This separation is based on the noise spectrum estimation described in the previous section. The estimated noise spectrum $\mid \widehat{N}(\omega) \mid^2$ and the observed spectral value $\mid Y(\omega) \mid^2$ are used to compute the signal-to-noise ratio :

$$SNR = 10 \log\left(\frac{\mid Y(\omega) \mid^2}{\mid \widehat{N}(\omega) \mid^2} - 1\right) \qquad (2)$$

This ratio is then compared to a threshold: if the SNR is greater than the threshold, the data is considered present; otherwise it is considered missing. Several values have been tested for the threshold, from -10 to +10. The best results were obtained with the threshold equal to zero. This value was used during all of the experiments presented in this paper. The masks resulting from this process are called estimated masks. We also computed masks directly from the speech and noise files, before they are added together. This method provides us with a priori masks, which are used to estimated the performance both of the missing data recognition and of the noise estimation independently.

Once the masks are obtained, the framework of the standard Gaussian mixture HMM can be modified to take into account the missing data [5]. In classical HMM systems, the emmitting probability of a state s_k is given by a Gaussian mixture probability density function :

$$f(x|s_k) = \sum_i w_{ki} N(x, \mu_{ki}, C_{ki}) \qquad (3)$$

where for each Gaussian i, w_{ki} is the weight of the Gaussian in the mixture, μ_{ki} is the mean vector, and C_{ki} is the covariance matrix. The components of the mean vector and of the covariance matrix corresponding to (x^p, x^m) can be separated in the same way that we separated the feature vectors:

$$\begin{aligned} \mu_{ki} &= (\mu_{ki}^p, \mu_{ki}^m) \\ C_{ki} &= \begin{bmatrix} C_{ki}^{pp} & C_{ki}^{mp} \\ C_{ki}^{pm} & C_{ki}^{mm} \end{bmatrix} \end{aligned}$$

The marginal pdf of the Gaussian mixture is then given by :

$$f(x_p|s_k) = \sum_i w_{ki} N(x^p, \mu_{ki}^p, C_{ki}^{pp}) \qquad (4)$$

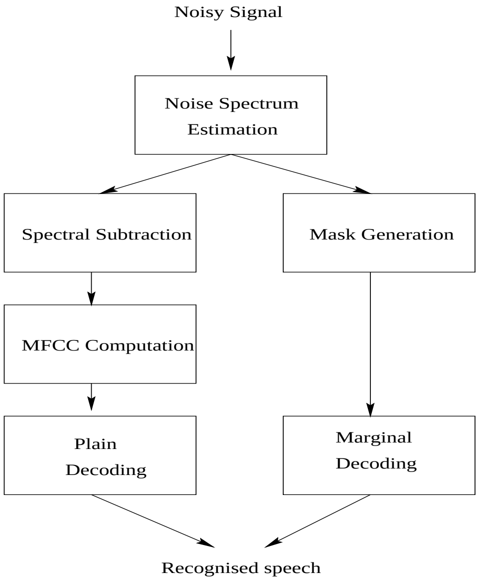

Figure 2: *Two different strategies*

Some constraints about the missing features can also be included in the computation of the estimated likelihood $f(x|s_k)$. These constraints take into account the fact that the missing spectral features can not be negative and can not exceed the noisy observed value. The estimated likelihood then takes the form :

$$\widehat{f}(x|s_k) = \sum_i w_{ki} N(x^p, \mu_{ki}^p, C_{ki}^{pp}) \int_0^{x_u} f(x_m|x_p, s_k) dx_m \qquad (5)$$

where $f(x_m|x_p, s_k)$ is the conditional pdf of the missing features x_m, given the present feature x_p and the state s_k. This method is called bounds-marginal decoding.

4. SPECTRAL SUBTRACTION

To deal with speech signals which are degraded by additive noise, speech enhancement techniques based on noise spectrum estimation have been used for a few decades[1]. If we consider a speech signal $s(t)$ which is degraded by additive noise $n(t)$, the resulting signal is then

$$y(t) = s(t) + n(t)$$

If we now consider $Y(\omega)$, the Fourier transforms of $y(t)$, and if we suppose that the noise $n(t)$ is uncorrelated with the speech signal $s(t)$, we can use the estimate of the noise energy obtained in section 2 to derive an estimate of the clean speech spectrum $\mid \widehat{S}(\omega) \mid^2$:

$$\mid \widehat{S}(\omega) \mid^2 = \mid Y(\omega) \mid^2 - \alpha \mid \widehat{N}(\omega) \mid^2 \qquad (6)$$

where α has to be optimized. Since this equation does not guarantee that $\mid \widehat{S}(\omega) \mid^2$ is positive, the negative values of $\mid \widehat{S}(\omega) \mid^2$ after subtraction are set to a constant, non-zero, minimum value.

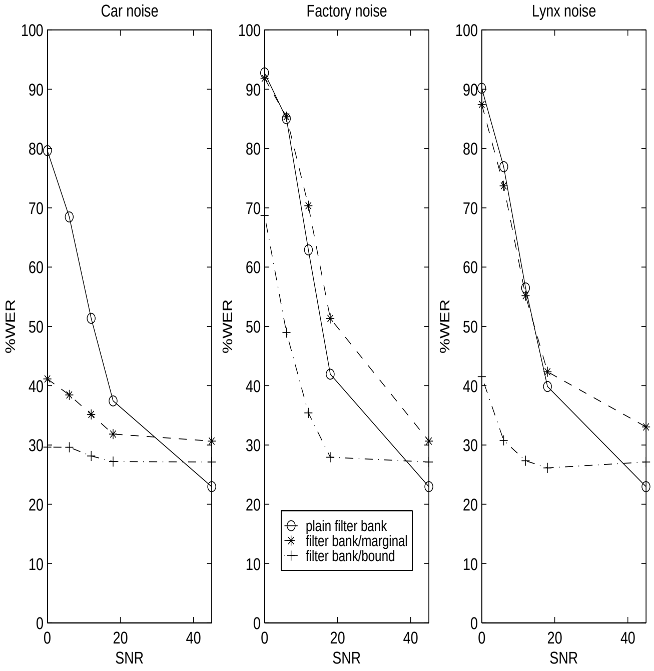

Figure 3: *Recognition error rate (%WER) vs. SNR for plain filter bank, marginal and bounds-marginal systems. Results are given on three kinds of noise*

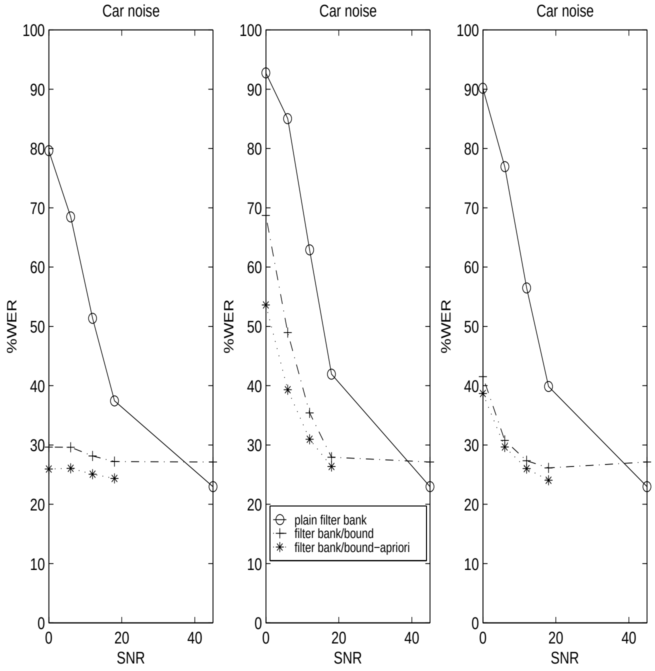

Figure 4: *Recognition error rate (%WER) vs. SNR for plain filter bank system, and bounds-marginal system with estimated and a priori masks. Results are given on three kinds of noise*

5. TESTS

The recognition system, based on Gaussian mixture HMMs, was trained with HTK [8] on the Numbers95 database - [3]. The system was composed of 81 triphones modeled by 3 state HMMs; each state had a 10 Gaussian mixture pdf and a diagonal covariance matrix. No language model was used. Numbers95 is a database of 31 different numbers obtained from continuous speech over the telephone. We used the standard training set of 3233 sentences and the standard test set of 1227 sentences.

During all of the missing data recognition experiments, we used 26 log mel-scaled filter bank coefficients, computed over a 32 ms hamming window, with a 10 ms shift. For the tests with spectral subtraction, we used 13 mel-cepstrum coefficients with first and second order derivative (for a total of 39 coefficients) since this parametrisation has been proved to be efficient.

The task was to recognise the 1227 test sentences from the Numbers95 databases under different noise conditions. We used three kinds of recorded noise from the NOISEX92 database [7]: car noise, factory noise,and lynx helicopter noise. The noise was added to clean speech at different SNR levels (18 db, 12 db, 6db, 0db).

6. RESULTS

The first set of experiments was designed to evaluate the recognition system based on SNR estimation with missing data recognition. Figure 3 presents the recognition results for the three kinds of noise at different SNR levels for both the filter-bank baseline system and the missing data system with marginal and bound methods. As show in the figure, the missing data system based on marginal decoding improved the baseline system only for the car noise at a low SNR level. The bounds-marginal method significantly improved the results of the filter bank baseline system for the three kinds of noise and at all noise levels. On average, this method gave an error rate reduction of 30% when compared to the baseline system. However, the bounds-marginal method was also more resistant to car noise. Note that both methods underperformed the baseline system for clean speech. This point is discussed in the next section.

The second set of experiments was designed to evaluate the accuracy of the SNR estimator. Indeed, the missing data recognition system relies highly on the SNR estimate, from which the masks are derived. In Figure 4, we compare the recognition performance of the bounds-marginal missing data system when used with either estimated or a priori masks. As expected, using a priori masks gave better results than using estimated masks for all noise conditions; but, on average, using a priori mask only yield 10% reduction of the error rate in comparison with using estimated masks.

Finally, the third set of experiments, illustrated in Figure 5, compared the missing data performance to the classical spectral subtraction system. The recognition system based on spectral subtraction outperformed the missing

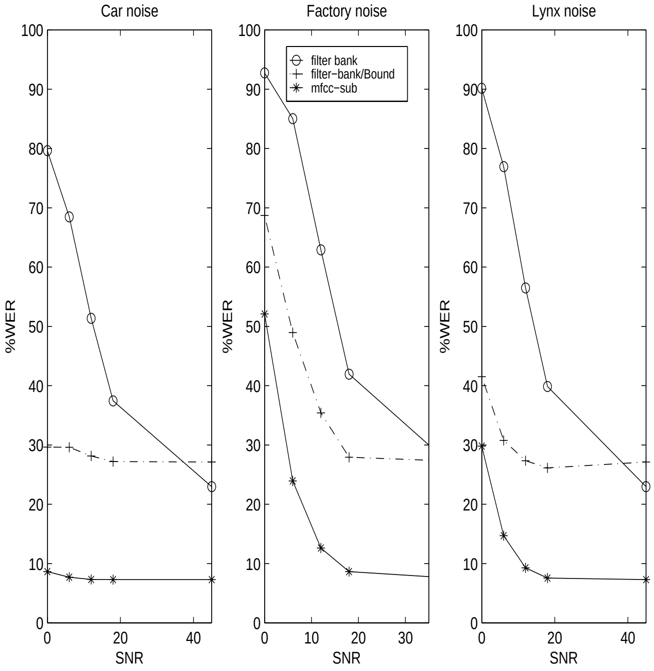

Figure 5: *Recognition error rate (%WER) vs. SNR for plain filter bank, bounds-marginal and MFCC with spectral subtraction systems. Results are given on three kinds of noise*

data system in every noise condition and on average, the error rate is 60% lower with spectral subtraction in comparison with the bounds-marginal system.

7. DISCUSSION

The experiments have shown that the missing data methods are more efficient in certain noise conditions (in our case, car noise). This may be due to the fact that the car noise is relatively band-limited whereas the lynx noise and the factory noise are spread over all the frequency band. In the latter case, the hypothesis that some parts of the feature vectors are free from noise is not satisfied. Therefore, there are not enough "clean" feature vector parts to perform a good recognition. However, when the noise is band limited, the bounds-marginal method showed good resistance to noise.

It was also shown that both methods underperformed the baseline system for clean speech. This problem is related to the choice of the threshold for the masks computation. The threshold of 0 dB was chosen for all of the experiments in order to maximise the performance under all noise conditions. Setting the threshold lower would have improved the recognition results in clean speech while degrading the performance in noise. This problem might be solved by using an adaptive threshold, based on the level of noise.

The main drawback of the missing data method is that it is based on non-orthogonal features, and, therefore, are outperformed by systems based on orthogonal features. This problem might be solved by using sub-band orthogonalization as it is usually done in multi-band systems [2].

8. CONCLUSION

In this paper we have compared two different strategies for reducing the effect of additive noise on speech recognition: missing data and spectral subtraction. The former has shown good resistance to noise under certain conditions but still needs to be improved in order to reach the performance of the latter. Some future directions for such improvements were discussed.

Acknowledgments:

This work is supported by the Swiss Federal Office for Education and Science (FOES) through the European TMR SPHEAR project.

9. REFERENCES

[1] Steven F. Boll. Suppression of acoustic noise in speech using spectral subtraction. *TransASSP*, ASSP-27(2), april 1979.

[2] Hervé Bourlard, Stéphane Dupont, and Christophe Ris. Multi-stream speech recognition. IDIAP-RR 7, IDIAP, 1996.

[3] 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.

[4] D. Mansour and H. Juang. The short time modified coherence representation and noisy speech recognition. *TransASSP*, 37(6):795–804, june 1989.

[5] A. C. Morris, M. P. Cooke, and P. D. Green. Some solutions to the missing features problem in data classification, with application to noise robust asr. *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, pages 737–740, 1998.

[6] A. P. Varga and R. K. Moore. Hidden Markov model decomposition of speech and noise. In *ICASSP*, pages 845–848, 1990.

[7] 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.

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