ELSEVIER

Speech Communication 34 (2001) 141–158

SPEECH COMMUNICATION

www.elsevier.nl/locate/specom

Assessing local noise level estimation methods: Application to noise robust ASR

Christophe Ris *, Stéphane Dupont

Faculté Polytechnique de Mons – TCTS, Multitel, Parc Initialis, Mons B-7000, Belgium

Abstract

In this paper, we assess and compare four methods for the local estimation of noise spectra, namely the *energy clustering*, the *Hirsch histograms*, the *weighted average method* and the *low-energy envelope tracking*. Moreover we introduce, for these four approaches, the *harmonic filtering* strategy, a new pre-processing technique, expected to better track fast modulations of the noise energy. The speech periodicity property is used to update the noise level estimate during voiced parts of speech, without explicit detection of voiced portions. Our evaluation is performed with six different kinds of noises (both artificial and real noises) added to clean speech. The best noise level estimation method is then applied to noise robust speech recognition based on techniques requiring a dynamic estimation of the noise spectra, namely spectral subtraction and missing data compensation. © 2001 Elsevier Science B.V. All rights reserved.

Zusammenfassung

Dans ce papier, nous nous proposons d'évaluer et de comparer différentes méthodes d'estimation locale du spectre de bruit: le *clustering* des énergies, les histogrammes de Hirsch, la méthode de la moyenne pondérée ainsi que le suivi de l'enveloppe de basse energie. Nous présentons également, en pré-traitement pour chacune de ces méthodes, le filtrage des harmoniques, une technique permettant de suivre plus efficacement des variations rapides de l'énergie du bruit. Le caractère harmonique de certaines portions de parole est exploité afin de réestimer le niveau de bruit pendant les périodes de sons voisés (sans détection explicite du caractère voisé ou non des segments de parole). Ces différentes approches ont été testées sur six types de bruit différents (artificiels et réels) ajoutés à de la parole claire. Les estimateurs de niveaux de bruit ainsi testés ont alors été appliqués à deux méthodes de reconnaissance de la parole robuste aux bruits basées sur la soustraction spectrale et la théorie des données manquantes. © 2001 Elsevier Science B.V. All rights reserved.

Keywords: Robust automatic speech recognition; Noise level estimation; Noise reduction; Spectral subtraction; Missing data

1. Introduction

Although current speech recognition systems tend to perform well even on difficult large vocabulary continuous speech tasks, their performance dramatically drops when they are used in adverse conditions such as background noise, reverberation or just strong mismatch between operating and training conditions. Speech recognition technology has now reached the level of maturity (both in terms of design and implementation) that could allow it to be used in everyday

* Corresponding author.
E-mail addresses: ris@tcts.fpms.ac.be (C. Ris), dupont@tcts.fpms.ac.be (S. Dupont).

PII: S0167-6393(00)00051-0

applications. Unfortunately, most of such applications potentially suffer from interference: voice dialing in cars, database access through mobiles, automatic information desks in shopping malls or train stations, ...

Many ideas are currently being developed within the speech community in order to improve the robustness of automatic speech recognition (ASR) systems, e.g. robust acoustic features (Hermansky and Morgan, 1994), spectral subtraction pre-processing techniques (Berouti et al., 1979; Boll, 1979), model-based compensation methods (Gales, 1997), missing data compensation (Cooke et al., 1997; Dupont, 1998), multi-band paradigm (Tibrewala and Hermansky, 1997; Mirghafori and Morgan, 1998). Most of these methods have in common the need for an estimation of the interfering noise level (or of the signal-to-noise (SNR) ratio) that should be local in time and in frequency, e.g. noise spectral magnitude estimation for spectral subtraction, selection of unreliable features in the time–frequency grid for missing data methods, estimation of a confidence level for sub-band recognizers in the multi-band approach, ...

In this paper, we address the problem of accurate local estimation of non-stationary noise level in order to be able to track relatively fast modulations of the noise spectral magnitude.

Noise spectral magnitude estimation is usually done by explicit detection of pure noise segments (speech pauses) (Korthauer, 1999; McKinley and Whipple, 1997; Sarikaya and Hansen, 1998). This can be very difficult in the case of varying background noise or if the SNR ratio is very low. In this paper, we evaluate four methods which do not need any explicit speech pause detection. We compare them to an "ideal" configuration where position of pure noise segments are known exactly. The first two methods are referred to as *energy clustering* (Bourlard et al., 1996) and *Hirsch histograms* (Hirsch and Ehrlicher, 1995), the third one is an *adaptive weighted average* estimation introduced by Hirsch and Ehrlicher (1995), and the last one, inspired from the work of Martin (1993), is based on *lower-energy envelope tracking*. In order to track more accurately quick changes in the noise spectrum, we also introduce the *harmonic filtering* strategy, a new pre-processing method for estimating the noise level during voiced parts of speech.

The evaluation of the different noise level estimation methods is performed with six different kinds of noise added to clean speech. Their performance is measured in terms of mean square error (MSE) between estimated values and actual noise.

The noise estimation method yielding the lowest overall MSE is then applied to noise robust ASR systems based on spectral subtraction or missing data compensation. These algorithms require good estimates of the noise spectral magnitude.

2. Methods for local noise level estimation

Different algorithms for local [1] noise level estimation have been tested. Our goal is to estimate the noise spectrogram from the noisy data. Hence, we have to discretize both the frequency axis and the time axis. The use of narrow frequency bands (200 Hz) leads to a good estimation of the noise magnitude spectrum while a frequent update of the estimations allows to closely match fast noise level changes.

All the methods that are investigated here are based on the following. First, the signal is analyzed using short-time spectra computed from short overlapping analysis frames (typically 32 ms windows with 16 ms overlap). Then, several consecutive frames are used in the computation of the local noise spectrum. The analysis frames that are used together form a *time segment*. The typical time span of this segment is several hundred milliseconds. Additionally, the methods are based on the following important assumptions:

- the time segments contain speech pauses or low-energy signal portions,
- the noise present in these time segments is more stationary than the speech itself.

The first assumption implies to use long segments so as to ensure the presence of speech pauses [2] or

[1] We mean local in time and in frequency.

[2] By "speech pause" we mean a non-speech portion, that is portions where only background noise is present.

low-energy signal portions, while the second assumption requires time segments short enough for the noise to be considered stationary. As a consequence, the length of the time segments will result from a trade-off between these two constraints. If the noise can be assumed to be stationary (e.g. from a priori knowledge of the noise conditions), we will prefer long time segments to satisfy the first assumption and hence to obtain more accurate estimates. If the noise is non-stationary, we will choose short segments so as to be able to track fast changes in the noise level.

The following sections describe the four noise level estimation methods that have been tested. For each algorithm, we give a pseudo-code which describes the processing done to obtain a noise level estimation from a limited time segment in a single frequency band.

2.1. Energy clustering

This method is based on the analysis of histograms of energy values within different frequency bands. These histograms can be used to estimate both the noise level and the speech level of the acoustic signal. Indeed, under the assumption that they are built on signal segments long enough to contain portions of speech together with speech pauses or low-energy signal portions, we can observe the following properties (see Figs. 1 and 2):

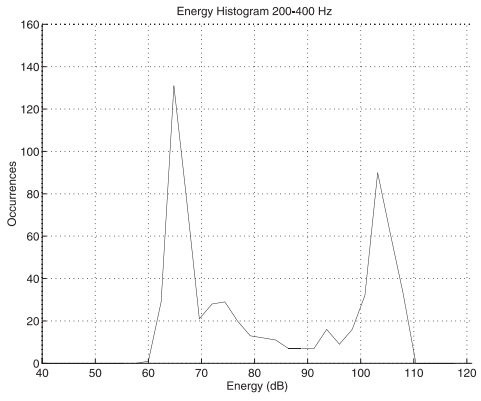

Fig. 1. Energy histogram for 6.5 s of clean speech (200–400 Hz frequency band).

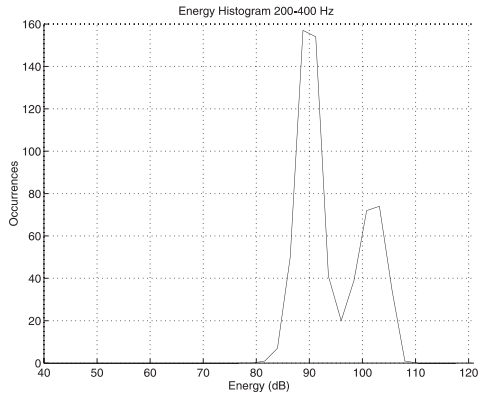

Fig. 2. Energy histogram for the same signal and same frequency band as Fig. 1 but with additive white noise at 0 dB SNR.

1. The histograms, whether they are computed on a noisy frequency band or not, contain basically two modes:
   (a) a low-energy mode related to the contribution of (possibly noisy) speech pause frames,
   (b) a high-energy mode related to the contribution of (possibly noisy) speech frames.
2. In general, the low-energy mode is higher and has a smaller variance than the high-energy mode. This is the case if, as commonly assumed, the energy of the silence (or noise) is more stationary than the energy of the speech signal.
3. The two modes are clearly separated in case of clean speech (see Fig. 1).
4. When noise is added in the observed frequency band, the two modes are getting closer (see Fig. 2) to eventually merge into one mode.

It is therefore possible to model the energy distribution in a particular frequency band by a two-mode model. We can, for instance, try to fit two Gaussian probability density functions (pdfs) to the histogram, as in (Van Compernolle, 1989). We can use the EM algorithm to tune the parameters of the Gaussian pdfs according to a maximum likelihood criterion.

An alternative to this method consists of using a two-centroid clustering algorithm (such as the *k-means* algorithm) in order to detect the two modes of the distribution. The lower centroid

value corresponds to the noise level while the higher centroid value corresponds to the noisy speech level. As a by-product, this method directly provides an estimate of the local SNR ratio, which can also be useful to many applications.

Note that both the clustering and the EM algorithms assume the presence of two modes. Therefore, they appear to be inaccurate when the two modes tend to merge, in other words, for low SNR values. This leads to an underestimation of the noise level for low SNR values. We propose a solution to this problem in Section 2.5.

The following pseudo-code (see Plate 1) describes the two-centroid clustering algorithm within a single frequency band. Note that the clustering is applied in the log-energy spectral domain.

2.2. Hirsch histograms

Just like the energy clustering method, this approach is based on the statistical analysis of the spectral energy envelope. It is derived from the Hirsch histograms method (Hirsch and Ehrlicher, 1995). Histograms of energy values are built for different frequency bands on signal segments of several hundred milliseconds. The principle is a consequence of the two assumptions stated above. Under these assumptions, the most frequent energy value (the histogram maximum) is related to the noise level in the current frequency band. However, this ideal behavior is not guaranteed, and in many cases, the most frequent energy value is much higher than the noise level. This is the case, for instance, for low-frequency bands where the speech signal has rather high energy levels. So, to avoid noise energy overestimation, the estimated noise level is limited to a (frequency dependent) dynamically updated threshold. This threshold is computed as the mean value of the energy minima in the signal segment multiplied by a factor larger than 1. The most frequent energy value under this threshold is considered as an estimate of the noise level. Moreover, the noise level is limited to the average energy of the signal segment in the current frequency band as no noise can occur with higher energy than the average energy inside the band (this is the case when there is only noise in the signal segment).

Note that we build the histograms in the log-energy spectral domain, whereas Hirsch formulation was in the magnitude spectral domain (Plate 2).

2.3. Weighted average

This method introduced by Hirsch and Ehrlicher (1995) is based on a first-order recursion and uses an adaptive threshold to stop the recursion when speech is most likely to be present. An estimate of the noise energy spectrum at time t is obtained by the first-order recursion,

$$N_t(\omega) = \begin{cases} \alpha N_{t-1}(\omega) + (1-\alpha) X_t(\omega) \\ \quad \text{if } X_t(\omega) \leqslant \beta N_{t-1}(\omega) \\ N_{t-1}(\omega) \\ \quad \text{otherwise,} \end{cases} \tag{1}$$

where $X_t(\omega)$ is the energy spectrum of the signal, $N_t(\omega)$ the estimated noise energy spectrum, $\beta = 2$ and $\alpha = 0.98$.

Recursion 1 tends to overestimate the noise in the frequency bands where the SNR is low. In order to reduce this error, a second order recursion is added,

$$\begin{aligned} \mathrm{var}_t(N(\omega)) = {} & \alpha \mathrm{var}_{t-1}(N(\omega)) \\ & + (1-\alpha)(X_t(\omega) - N_t(\omega))^2. \end{aligned} \tag{2}$$

Eqs. (1) and (2) estimate the noise mean and variance, respectively. These recursions are applied only if

$$|X_t(\omega) - N_{t-1}(\omega)| \leqslant k \sigma_N(\omega), \tag{3}$$

where $\sigma_N(\omega)$ is the frequency-dependent noise standard deviation and k is a tunable threshold parameter.

We implemented a variation of this method. The recursions are applied using estimations of the noise energy and of the noise energy variance. Unlike the previous description, however, these estimations are obtained by analyzing a time segment of several frames. Note also that the formulation in Hirsch and Ehrlicher (1995) was in the magnitude spectral domain (Plate 3).

Plate 1

Initialization of the two cluster centroids : `low_centroid` *for low energies and* `high_centroid` *for high energies. Average signal segment energy is taken as a reference for initialization.*

```
low_centroid = energy_average * 0.999;
high_centroid = energy_average * 1.001;
```

K-means clustering iterations

```
do {
  new_low_centroid = new_high_centroid = 0.0;

  foreach (signal_energy) in (signal_segment)
  {
```

computes distance between signal energy and centroids

```
    low_distance = (low_centroid - signal_energy)^2;
    high_distance = (high_centroid - signal_energy)^2;
```

assign current frame energy to the closest centroid

```
    if (low_distance <= high_distance)
    {
      new_low_centroid += signal_energy;
      low_count++;
    }
    else
    {
      new_high_centroid += signal_energy;
      high_count++;
    }
  }
```

`new_low_energy` *and* `new_high_energy` *are the new centroid positions.*

```
  new_low_centroid /= low_count;
  new_high_centroid /= high_count;
```

global distortion is used to stop K-means iterations.

```
  distortion = ((new_low_centroid - low_centroid)^2
               + (new_high_centroid - high_centroid)^2)^(1/2);
```

assign updated centroid positions to `low_centroid` *and* `high_centroid`

```
  low_centroid = new_low_centroid;
  high_centroid = new_high_centroid;
```

use distortion threshold to stop K-means iterations.

```
} while (distortion > DISTORTION_THRESHOLD);
```

the lower centroid is used as noise level estimate.

```
noise_estimate = low_centroid;
```

Plate 2

Prepare the histogram bins. `min_energy` *is the lowest energy in the signal segment, while* `max_energy` *is the average segment energy.* `HISTO_BIN_COUNT` *is the number of containers (bins) in the histogram.* `energy[bin]` *is the container center.*

```
float step_size = (max_energy - min_energy) / HISTO_BIN_COUNT;
foreach (bin) in (histogram)
{
  energy[bin] = min_energy + step_size * (bin + 0.5);
}
```

build histogram with signal energies lower than a dynamically updated threshold (see below). `histo[bin]` *is the number of elements in the bin.*

```
foreach (signal_energy) in (signal_segment)
{
  if (signal_energy < min(max_energy, threshold))
  {
    bin = (signal_energy - min_energy) / step_size;
    histo[bin]++;
  }
}
```

for each bin of the histogram, compute the mean number of elements within a group of `span` *containers around it.* `span` *is a function of the bin center. The center of the bin having the maximum value is the estimation of the noise level. This allows to smooth the histogram for the noise level estimation in the cases of low SNR.*

```
peak_value = -INF;
foreach (bin) in (histogram)
{
  if ( histo[bin] > peak_value)
  {
    peak_value = histo[bin];
    peak_position = bin;
    noise_estimate = energy[peak_position];
  }
}
```

update the energy threshold.

```
threshold = noise_estimate + 9dB
```

2.4. Low-energy envelope tracking

The method described in this section is inspired from the work of Martin (1993) and is based on automatically tracking the low-energy envelope of the signal within frequency bands. The energy minima correspond to the portions of signal where no speech occurs, that is, to frames containing only noise. The average value of these minima is used as an estimate of the noise level in the current frequency band. This method is based on the assumption that the noise is more stationary than the speech. Therefore, the mode corresponding to the noise has a very low variance and its mean can be estimated from the mean of the energy minima.

Practically, the algorithm can be described as follows: for each frequency band, and on the basis of a time segment of N frames, the n (typically $n = 0.2N$) frames with the lowest energies are averaged to estimate the noise level within the considered frequency band.

This method generally leads to an underestimation of the noise level. This is due to the fact that we only consider the energy minima in the current signal segment, and that the noise variance is not negligible. In Section 2.5, it will be shown how it is possible to compensate for this effect (Plate 4).

2.5. Correction of the measures

The methods described above can lead to gross estimation errors, especially for extreme SNR

Plate 3

Mean (`mean`*) and variance (*`var`*) are computed with signal energies for the current signal segment satisfying criterion of equation (3). They are used to update the noise level estimation and the noise variance calculated so far (*`noise_estimate`* and* `variance`*) according to equations 1 and 2*

```
count = 0;
mean = 0;
var = 0;
foreach (signal_energy) in (signal_segment)
{
  tmp = signal_energy - noise_estimate;

  if (fabs(tmp) <= k * (variance)^(1/2))
  {
    mean += signal_energy;
    var += (signal_energy)^2;
    count++;
  }
}

mean /= count;
var = var/count - (mean)^2;

noise_estimate = alpha * noise_estimate + (1 - alpha) * mean;
variance = alpha * variance + (1 - alpha) * var;
```

Plate 4

Find the `n` *lowest energies in the signal segment.* `signal_energies` *is the vector of energies on the speech time segment.*

```
sorted_energies = sort(signal_energies);
noise_estimate = average(sorted_energies[1 -> n]);
```

values. To compensate for these undesired effects, we define for each method, a correction mapping depending on the estimated SNR.

We add a known amplitude-modulated white noise to clean speech training data, for a wide range of SNR values (from 0 dB to 30 dB). Estimated noise energy spectra $\hat{N}_t(\omega)$ are then obtained on this noisy training set (using time segments). The speech plus noise energy spectra $X_t(\omega)$ are also estimated using the same time segments. For the energy clustering approach, the high-energy cluster is used as speech plus noise energy estimation. For the three other methods, $X_t(\omega)$ (within each frequency band) is calculated as the mean energy of the frames whose energy is above the mean energy within the whole time segment.

These estimations, together with the actual noise energy spectra $N_t(\omega)$, are used to define a mapping from the estimated noise energy to the real noise energy. The SNR-dependent [3] correction factor of this mapping is the mean (across time and frequency) of the estimation errors $N_t(\omega) - \hat{N}_t(\omega)$.

Note also that, whatever the method, the noise level estimations are smoothed using a low-pass filter, in order to avoid potential spikes.

[3] The SNR is computed using estimates $X_t(\omega)$ and $\hat{N}_t(\omega)$.

2.6. Harmonic filtering

Previously described methods give noise level estimates from speech segments. These segments have to be long enough to contain "silent" frames leading to a low-energy mode in the energy histograms of the different frequency bands. Continuous speech segments hopefully contain such "silent" frames:

- during the closures,
- during unvoiced fricatives at low frequencies: these kinds of sounds have very low energy under 2 kHz,
- during vowels at high frequencies.

We also know that vowels are quasi-periodic sounds leading to short-term spectra composed of harmonics superposed to the noise spectrum. At inter-harmonic frequencies, the signal energy is mainly due to noise (the effect of the shape and length of the analysis window used to compute the short-term spectra will be discussed later in this section). At harmonic frequencies, both signals contribute to the energy. Energy values of the narrow-band spectra within a limited frequency band can be used to build energy histograms similar to those of the previous sections. If we can assume that the noise is white within the considered narrow frequency band, the histograms should contain two modes:

- a low-energy mode for energy values between the harmonics, essentially related to noise,
- a high-energy mode for energy values near the harmonics, related to speech with superimposed noise.

The assumption about the noise shape is valid for wide-band noises within narrow frequency bands.

Speech segments without vowels unfortunately lead to mono-modal histograms and do not allow noise level estimation. As before, we can have recourse to time segments. This time, these segments are required to contain either silence portions, or else vowel sounds. Consequently, they could be shorter than the segments used with the previous methods. From these assumptions, we can devise a *harmonic filtering* method where regions of a narrow-band spectrogram are used as input to the previous noise level estimation methods for further processing. This approach should allow to shorten the time segments, and hence to increase the estimation accuracy in the case of non-stationary noises. [4]

We have been using 200 Hz frequency bands (centered around multiples of 100 Hz) with 100 Hz overlap. They are narrow enough to match colored noises and wide enough to include spectral minima, knowing that the initial windowing leads to the convolution of the "ideal" harmonic spectrum with a function whose main lobe is relatively wide. We have been using the Hanning window (Hamming could also be used) as it shows a good compromise between the main lobe width and the secondary lobe attenuation. The width of the first lobe is $4/d$ with d being the length (in s) of the window. For a 64 ms long window (512 points at a 8 kHz sampling rate), the first lobe is 62.5 Hz wide. As the pitch of vocal sounds is generally higher than 100 Hz, the harmonics are clearly separated with an analysis based on 64 ms windows. We could go down to 40 ms but this would anyway lead to the computation of a 512 points FFT. Let us emphasize that for some low-pitched male voices, [5] a typical analysis window of 30 ms is too short to get a spectrum with harmonics clearly separated by "spectral valleys". The secondary lobe attenuation also depends on the window length. For a 64 ms window, the first secondary lobe is 29 dB under the main lobe. When the noise level is very low compared to the speech level, this method may lead to noise level overestimation. At this level however (-29 dB), the noise should not affect the speech recognition process. As illustrated in Fig. 3 with the classical methods described earlier, a time vector of several hundred milliseconds containing the average energies in limited (200 Hz wide) frequency bands is used for the noise level estimation. In the case of harmonic filtering, as the analysis is based on narrow-band spectra, inter-harmonic energies can participate in

[4] Remembering that increasing the length of the analysis segments allows to increase the probability of including a decent number of "silent" frames, at the expense of worse results with non-stationary noises.

[5] As well as for "vocal fry" (doubling of the signal period) portions.

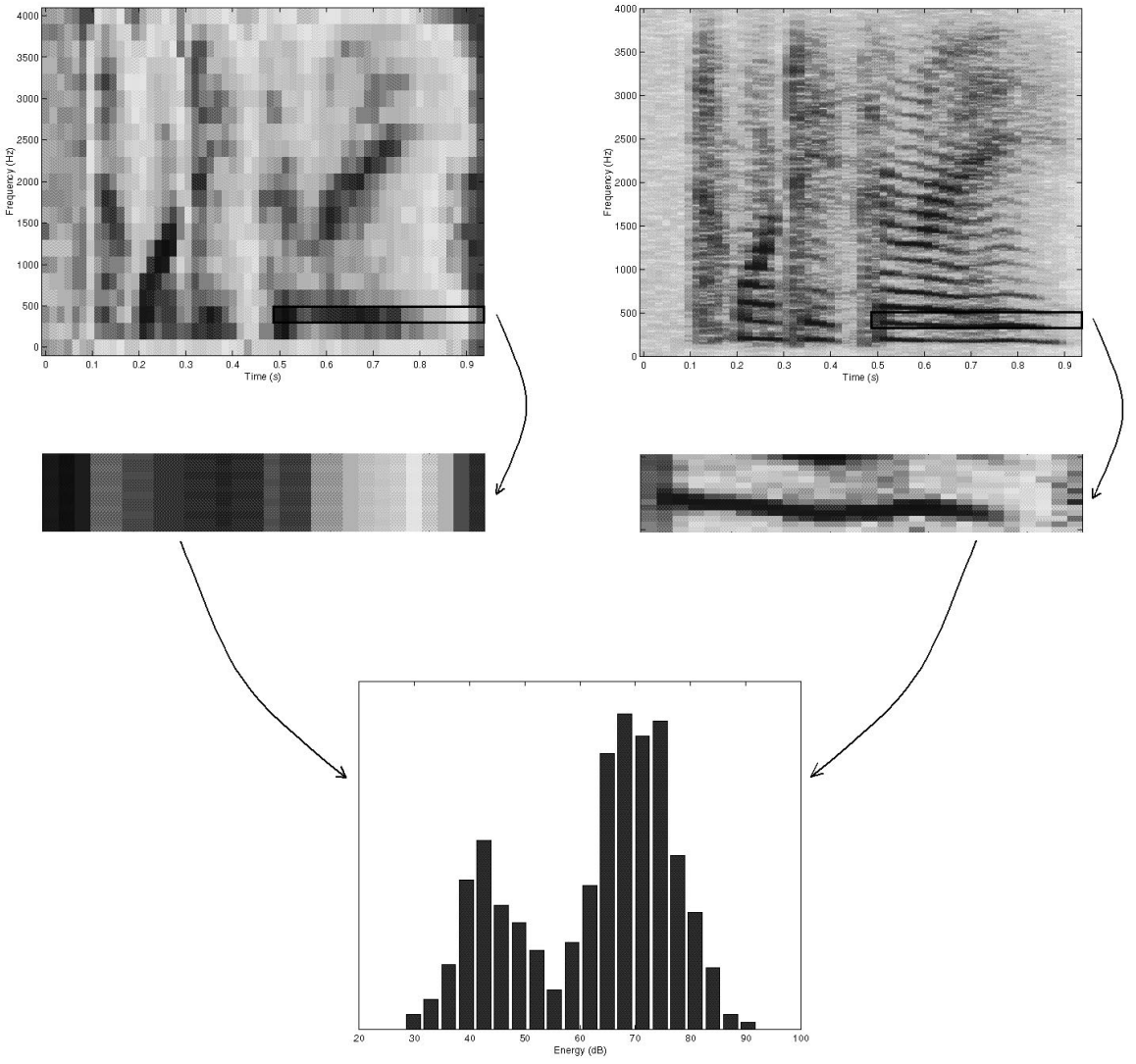

Fig. 3. Illustration of the harmonic filtering method. On the left, the classical approach based on the wide band spectrogram. On the right, the harmonic filtering method based on the narrow-band spectrogram. With this method, inter-harmonic energies participate in the noise estimation (illustrated by the energy histogram in this figure).

the estimation. Harmonic filtering is implemented by using a time–frequency matrix of energies from the narrow-band spectrogram. This matrix spans several hundred milliseconds and its width is 200 Hz. It is used as input to the corresponding frequency channel of one of the four noise level estimation methods described earlier.

2.7. Hybrid approach

The strategy described in the previous section (*harmonic filtering*) does not allow us to measure the level of sine waves or periodic stationary noises. Their harmonics are filtered out just like the harmonics of voiced speech sounds.

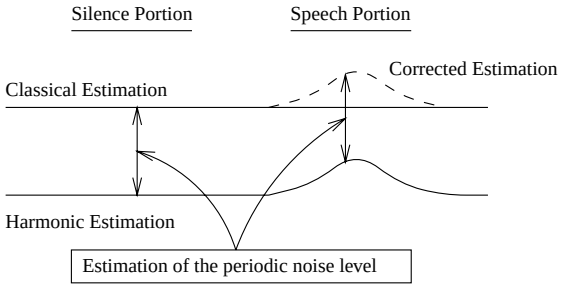

Fig. 4. Hybrid estimation method.

The classical [6] methods described in Sections 2.1–2.4 allow us to measure the energy of periodic noises but need longer temporal segments, with the assumption that the noise is stationary on longer signal portions. Periodic noises can however be superimposed on wide-band noises.

To benefit from both approaches, we propose to combine their results by the following method. During signal portions labeled as "speech pause" [7] by the classical method, the presence of a periodic noise will lead to a noise level estimation which is higher for the classical method than for the harmonic filtering method, because the contribution of harmonic noise is not present when harmonic filtering is used. The difference between the two measures is an estimate of the periodic noise level within the considered frequency band. Assuming that the periodic noise is stationary, it is then possible to correct the output of the harmonic filtering method using this estimate. As illustrated in Fig. 4, a small increase in the wide band noise level will not be detected by the classical method and the corrected harmonic filtering method will give better results.

2.8. Speech/silence detection

Other noise level estimation procedures rely on explicit speech/silence detection (Korthauer, 1999; McKinley and Whipple, 1997; Sarikaya and Hansen, 1998). As an upper bound for the performance of this class of methods, we used an artificial speech/silence detector based on a forced HMM-based speech/silence alignment of the clean version of the speech data. The estimated noise spectrum is updated during silence portions and linearly interpolated during speech portions. This method will be used as a reference.

2.9. Statistics

Observation of a speech spectrogram shows that:

1. The energy of unvoiced fricatives is very low at low frequencies.
2. The energy of vowels is low at high frequencies.

These "silent" portions introduce a low-energy mode in the energy histograms and allow to estimate the noise level during continuous speech utterances, using one of the aforementioned methods. Moreover, they suggest that the optimal duration of the speech segments could depend on frequency. In order to get some insight into this assumption, we collected statistics on the training data of the Resource Management (Price et al., 1988) corpus: this database allows to analyze the 0–8000 Hz frequency range. An analysis based on frequency bands each spanning 200 Hz (with 100 Hz overlap) was performed using 64 ms frames with 32 ms overlap. For different speech segment lengths, we estimate the probability that the segment contains at least a pre-defined proportion of "silent" frames. For the classical methods (Sections 2.1–2.4), the following elements are identified as silent:

1. portions labeled as silence according to the forced Viterbi alignment of an HMM model,
2. portions whose energy is under a threshold which is 18 dB below the mean speech energy of the considered band. These are essentially related to unvoiced fricatives at low frequencies and to vowels at high frequencies.

For the versions of the estimation methods using harmonic filtering, the frames labeled as vowels according to a forced Viterbi alignment are also identified as silent. These silent frames are most likely related to the lower mode of the histograms

[6] By classical we mean that no harmonic filtering is applied to the four methods described in this paper.

[7] Speech pause portions can be detected as a by-product of the noise level estimation procedure, using a threshold on the signal energy.

discussed previously, allowing for noise level estimation.

Results for the classical methods (that is no harmonic filtering) are presented in Figs. 5 and 6. The first figure plots the probability that a segment of x frames contains more than 20% of silence; the different curves are for different frequency bands, centered around multiples of 500 Hz. The second figure plots this probability for the different frequency bands (in abscissa); the segment length goes from 200 ms to 2000 ms, with 200 ms steps. The lowest and top curves are respectively for the 200 ms and 2000 ms conditions. For a given probability, we can observe in Fig. 5 an important variation in the segment length: from 200 ms to 1000 ms for a 0.9 probability. This indicates that the speech segment can be shorter for some frequency bands than for others.

At low frequencies, unvoiced fricatives are silent and at high frequencies, some vowels are silent. At medium frequencies, neither the fricatives nor the vowels have a low energy and only the silence frames are really silent. This explains the "U" shape of the curves in Fig. 6. Indeed, the probability that a time segment contains more than 20% of silence decreases if the number of silence frames is decreased. Note that the curves drop in the 0–500 Hz range. This is due to the low value of the mean speech energy at those frequencies.

These experiments were reproduced assuming that the harmonic filtering versions of the estimation methods remove the vowel sounds perfectly. The results are presented in Figs. 7 and 8. We can observe an important decrease in the minimum segment length: 300 ms segments will suffice to get at least 20% of silence for 90% of the database. The probability of having more than 20% of silent

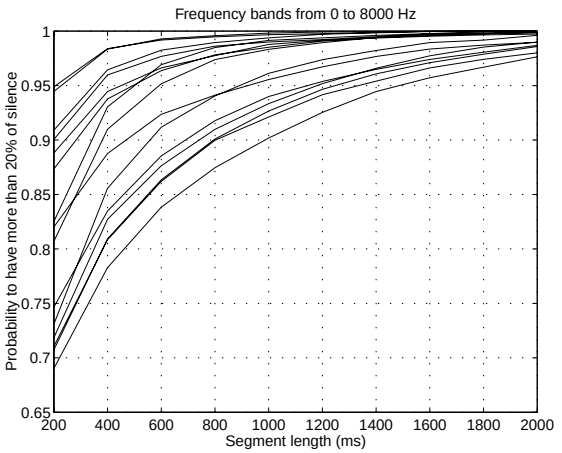

Fig. 5. Probability of having more than 20% silence for frequency bands centered around multiples of 500 Hz, each spanning 200 Hz (classical methods).

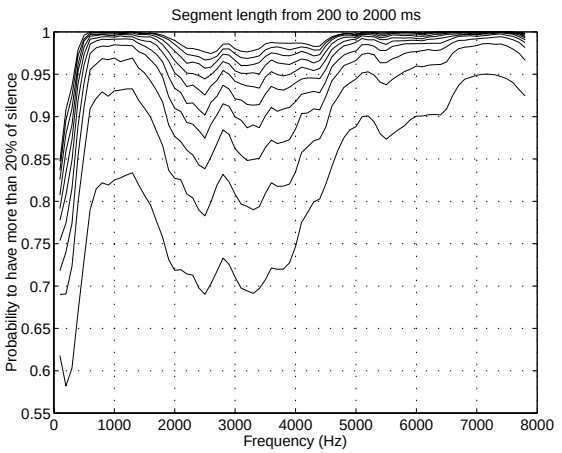

Fig. 6. Probability of having more than 20% silence for 10 time segment lengths ranging from 200 (lowest curve) to 2000 ms (classical methods).

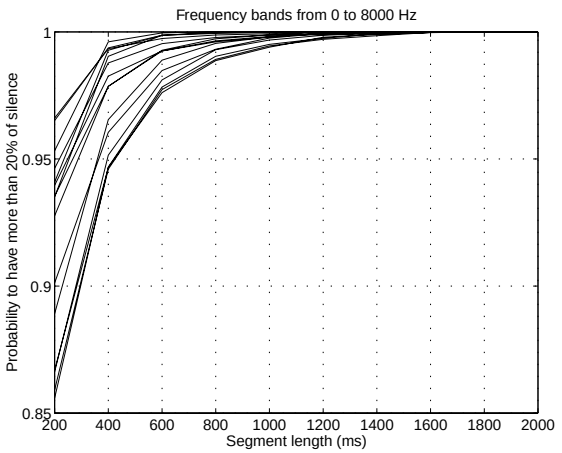

Fig. 7. Probability of having more than 20% silence for frequency bands centered around multiples of 500 Hz, each spanning 200 Hz (methods using harmonic filtering).

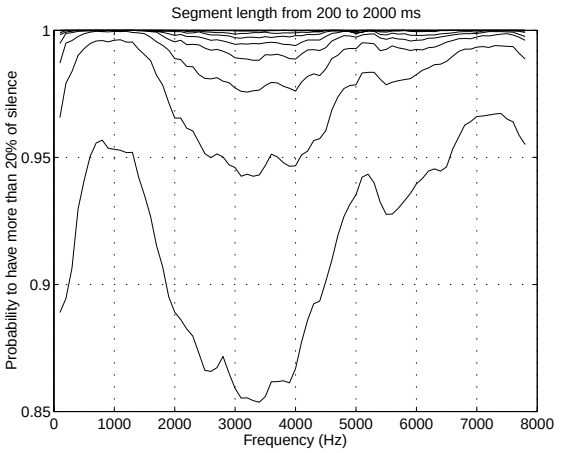

Fig. 8. Probability of having more than 20% silence for 10 time segment lengths ranging from 200 ms (lowest curve) to 2000 ms (methods using harmonic filtering).

frames is high at low frequencies and decreases to reach a minimum around 3000–3500 Hz. It finally increases at higher frequencies.

These statistics a posteriori justify the order of magnitude of the segment lengths. They could be used to optimize the segment length with the frequency band. This problem is still under investigation.

3. Comparison and assessment of the local noise estimation methods

The noise estimation algorithms and the effectiveness of harmonic filtering are tested in two different ways. In this section, we study the deviation of the estimated noise level from the actual noise level. These experiments are useful to identify the conditions for which the estimation methods and the harmonic filtering may be expected to be effective. In the next section, we will apply the method yielding the best noise estimates to speech recognition systems using robust strategies. The experiments intend to test whether "accurate" noise estimates can contribute to an improvement of ASR system performance.

The four methods described earlier (energy clustering, Hirsch histograms, weighted average and low-energy envelope tracking) were compared, without and with harmonic filtering. In addition, we used the ideal speech/silence detection algorithm as reference estimator. In order to obtain a fair comparison, all these methods have been implemented in the same program, sharing as many routines as possible (speech analysis, data buffering, ...). The program has been designed to be independent of any further processing (non-linear frequency scaling, frame rate, magnitude compression, filtering, ...) that could be applied to speech data. We process noisy samples and synthesize a time signal which is an image of the estimated noise synchronized with the noisy signal.

For all the experiments, we have been using 200 Hz frequency bands with 100 Hz overlap. Three different values for the time segment were compared: 250, 500 and 750 ms. When we did not use harmonic filtering, the analysis frames were 32 ms long (with 16 ms overlap). In all the experiments where we used harmonic filtering however, the analysis frames were 64 ms long, for the reason exposed earlier.

We used the following experimental set-up. We used connected digit string utterances from the Numbers'93 database (200 digit strings or 336 s of speech). Six kinds of noises were added to the clean speech:

- white Gaussian noise with amplitude modulation (SNR $= 15 \pm 15$ dB) at 0.5 Hz ($N1$),
- white Gaussian noise with amplitude modulation (SNR $= 15 \pm 15$ dB) at 1.0 Hz ($N2$),
- moving car noise (*2cars001* file) from the *Madras* (ULg, 1998) database ($N3$),
- helicopter noise from the *Noisex* (Varga et al., 1992) database ($N4$),
- factory noise from the *Noisex* database ($N5$),
- in-car noise from the *Noisex* database ($N6$).

The signal from the *Madras* and *Noisex* databases were first downsampled to 8 kHz.

The first three types of noises are highly non-stationary. Noise $N2$ is the least stationary of the six noise types, $N3$ has an amplitude-modulation frequency of $\sim$0.5 Hz . Noise $N4$ contains a strong harmonic structure. Noise $N5$ is a combination of a stationary background noise and unpredictable impulsive noises (hammer blows, siren wails, ...).

For each kind of noise, speech and noise were added such that the mean energy of each utterance

is 15 dB above the mean energy of the whole noise signal. Evaluation results for the nine local noise level estimation techniques are summarized in Table 1 for a frequency band covering the second formant (700 Hz→1600 Hz). The noise level estimations from the 200-Hz frequency bands spanning the 700–1600 Hz region are added to obtain a noise level estimation for the desired region. Deviation is measured as the MSE between this estimation in dB and the real noise level for the 700–1600 Hz region. Results are therefore expressed in dB^2. Results obtained with the artificial speech/silence detection are globally the best. Note that for the particular case of the least stationary noise ($N2$), the automatic method performs even better than the "ideal" speech/non-speech detection algorithm.

For the real noise types ($N3$–$N6$), long time segments (750 ms) give better performance. For $N1$, a lower error is obtained using 500 ms segments. For $N2$, the least stationary of the six kinds of noise, even shorter (250 ms) time segments would be preferred. As can be seen, harmonic filtering generally yields better results. Improvements are mainly noticeable in the case of short (250 ms) time segments. In the case of noise $N4$ (helicopter noise) however, harmonic filtering hurts a little. This is due to the fact that the noise spectra contain harmonic peaks that are also filtered out by this method thus leading to

Table 1
Noise level estimation on six kinds of noise[a]

Method	Harmonic filtering	Time segment length	$N1$	$N2$	$N3$	$N4$	$N5$	$N6$
Energy clustering		250	70.3	77.7	27.1	7.3	30.3	65.2
	No	500	43.8	84.1	16.3	7.2	16.5	51.1
		750	52.0	122.9	14.1	11.3	14.6	48.4
		250	41.2	52.0	11.7	8.7	12.0	11.4
	Yes	500	32.0	75.8	9.1	19.7	9.6	8.1
		750	48.9	126.2	9.7	20.4	9.5	7.3
Hirsch histograms		250	41.4	55.9	14.2	4.7	16.2	29.3
	No	500	34.1	95.6	8.5	3.7	9.9	19.3
		750	54.6	155.7	8.6	3.8	10.1	16.9
		250	35.0	46.6	9.7	16.2	11.4	9.5
	Yes	500	34.8	73.8	8.0	10.2	10.6	7.4
		750	45.8	117.8	8.2	9.0	10.7	6.9
Weighted average		250	50.6	77.8	22.0	9.3	24.8	69.7
	No	500	46.2	103.8	14.8	7.5	16.8	61.2
		750	55.7	150.2	12.7	7.1	14.9	58.7
		250	37.7	73.2	11.1	36.3	10.3	6.5
	Yes	500	35.8	86.5	11.6	26.3	9.7	6.1
		750	48.9	121.3	12.1	26.4	10.2	6.1
Low-energy envelope tracking		250	55.1	63.8	20.4	5.7	22.3	51.5
	No	500	29.4	76.0	7.4	3.8	8.4	32.5
		750	47.8	131.0	7.2	3.8	8.1	29.7
		250	39.2	50.1	10.7	7.7	12.9	13.1
	Yes	500	27.4	68.8	6.4	9.1	8.7	7.8
		750	40.1	121.6	6.3	9.5	8.3	6.7
Speech/silence detection			16.9	99.5	4.7	2.7	7.1	2.7

[a] Evaluation results for nine local noise level estimation techniques for a frequency band covering the second formant (700–1600 Hz). The noise level estimations from 200-Hz frequency bands spanning the 700–1600 Hz region are added to obtain a noise level estimation for the desired region. Deviation is measured as the MSE between this estimation in dB and the real noise level for the 700–1600 Hz region. Results are therefore expressed in dB^2.

Table 2
Noise $N4$ – use of hybrid approach in order to improve the noise level estimation in case of harmonic noise

Method	Harmonic filtering	Segment length	$N4$
Energy clustering	Hybrid	250	5.9
		500	4.7
		750	4.9

underestimation of the noise level. The hybrid approach can be used to solve this problem. For this experiment, we combined classic noise estimation and harmonic filtering as introduced in Section 2.7. Applied to the energy clustering method, this technique leads to a significant improvement of the estimation accuracy as can be seen in Table 2.

In this section, we have seen how harmonic filtering can be used to improve the local noise level estimation. The use of local noise estimates in a noise robust ASR set-up will be discussed in the next section.

4. Application to automatic speech recognition

4.1. Spectral subtraction

Spectral subtraction follows from the speech and noise independence assumption. As a result, an estimation of the speech spectrum is obtained by subtracting the noise energy spectrum from the signal energy spectrum. Generally, spectral subtraction introduces time-varying peaks and valleys in the energy spectrum. These are perceived as an annoying "musical" noise and could have a significant impact on the resulting performance of a speech recognition system.

To reduce the effect of this noise, Berouti et al. (1979) proposed a method where an overestimation (α parameter) of the noise is subtracted from the corrupted signal, and where valleys are filled with a fraction (β parameter) of the noise energy spectrum. Hence, energy spectral subtraction is implemented as follows:

$$E(\omega) = \begin{cases} \hat{S}(\omega) & \text{if } \hat{S}(\omega) > \beta N(\omega), \\ \beta N(\omega) & \text{otherwise,} \end{cases} \tag{4}$$

with

$$\hat{S}(\omega) = X(\omega) - \alpha N(\omega), \tag{5}$$

and

$$\alpha \geqslant 1 \text{ and } 0 < \beta \ll 1, \tag{6}$$

where $X(\omega)$ is the corrupted short-term input energy spectrum, $N(\omega)$ the estimation of the noise energy spectrum, and $E(\omega)$ is the enhanced short-term energy spectrum. The α parameter is the overestimation factor and β is called the spectral floor.

According to Berouti et al. (1979), distorsions of the speech signal can further be reduced using an SNR-adaptive α,

$$\alpha = \alpha_0 - \mathrm{SNR}/s \quad \text{for } -5 \text{ dB} \leqslant \mathrm{SNR} \leqslant 20 \text{ dB}, \tag{7}$$

$$\alpha = \alpha_0 + 5/s \quad \text{for } \mathrm{SNR} \leqslant -5 \text{ dB}, \tag{8}$$

$$\alpha = 1 \quad \text{for } \mathrm{SNR} \geqslant 20 \text{ dB}, \tag{9}$$

where α_0 is the desired value of α for $\mathrm{SNR} = 0$ dB, and s is chosen to have $\alpha = 1$ when $\mathrm{SNR} = 20$ dB. SNR is a function of the local SNR ratio:

$$\mathrm{SNR} = 10 \log_{10} \left(\frac{X(\omega)}{N(\omega)} \right). \tag{10}$$

4.2. Missing data

Application of missing data techniques in robust ASR (Cooke et al., 1999; Vizinho et al., 1999) requires to solve two subproblems: the identification of reliable spectro-temporal regions and recognition techniques to handle incomplete data. In this paper, we apply the SNR criterion (Cooke et al., 1999) to the first problem. The missing data imputation is used to solve the problem of reconstruction of missing data from the reliable data (Dupont, 1998).

4.2.1. Identification of the missing components

The estimate of the noise energy spectrum $N(\omega)$ can be used to identify the regions dominated by noise. This SNR criterion treats data as unreliable when the estimated SNR is negative,

$$\mathrm{SNR} = 10\log_{10}(S(\omega)/N(\omega)) < 0 \rightarrow S(\omega) < N(\omega), \quad (11)$$

where $S(\omega)$ is the clean signal energy spectrum, and assuming signal and noise are decorrelated, $X(\omega) = S(\omega) + N(\omega)$ is the energy spectrum of the noisy signal. Adding $N(\omega)$ to both sides of (11), the SNR criterion becomes

$$S(\omega) + N(\omega) < 2N(\omega) \rightarrow X(\omega) < 2N(\omega). \quad (12)$$

Regions where the noisy signal energy is less than 3 dB above the noise energy are identified as missing.

Generalized spectral subtraction could also be used to derive identification criteria, as in (El-Maliki et al., 1998).

4.2.2. Data reconstruction

Observation vectors X are assumed independent and identically distributed according to a probability density function made of K multi-dimensional Gaussian distributions. The ith distribution is characterized by the following parameters: w^i, the distribution weight, μ^i, the distribution mean and C^i, its covariance matrix. Some elements of X are labeled as missing and X can then be reorganized as follows:

$$X = (X_\mathrm{p} X_\mathrm{m}), \quad (13)$$

X_p for the present components and X_m for the missing components. In a similar way, we can reorganize the elements of the mean vectors and covariance matrices characterizing the probability density function of X,

$$\mu^i = (\mu^i_\mathrm{p} \mu^i_\mathrm{m}), \quad C^i = \begin{bmatrix} C^i_{\mathrm{pp}} C^i_{\mathrm{pm}} \\ C^i_{\mathrm{mp}} C^i_{\mathrm{mm}} \end{bmatrix}. \quad (14)$$

We would like to reconstruct a complete vector solely on the basis of present components. Reconstruction will be done using the conditional distribution of missing components according to present components. This distribution is of the Gaussian form. The reconstructed elements will be the mean of this distribution, that is, for the ith Gaussian,

$$X^i_{\mathrm{m|p}} = \mu^i_\mathrm{m} + (C^i_{\mathrm{pm}})^t (C^i_{\mathrm{pp}})^{-1} (X_\mathrm{p} - \mu^i_\mathrm{p}). \quad (15)$$

Considering the multi-Gaussian distribution, the reconstructed value is computed as follows:

$$X_{\mathrm{m|p}} = \frac{\sum_{i=1}^K w^i \phi(X_\mathrm{p}, \mu^i_\mathrm{p}, C^i_{\mathrm{pp}}) X^i_{\mathrm{m|p}}}{\sum_{i=1}^K w^i \phi(X_\mathrm{p}, \mu^i_\mathrm{p}, C^i_{\mathrm{pp}})}, \quad (16)$$

where w^i is the weight for the ith distribution and $\phi(X_\mathrm{p}, \mu^i_\mathrm{p}, C^i_{\mathrm{pp}})$ is the associated multi-dimensional Gaussian distribution. This term allows to weigh the contributions of the different Gaussians according to the position of the present data vector in the parameter space. Note that covariance matrices are often reduced to diagonal matrices, although this simplification does not allow to model the correlation between the feature vector elements, which could be important between adjacent frequency channels. In this case, Eq. (15) simply becomes $X^i_{\mathrm{m|p}} = \mu^i_\mathrm{m}$.

An alternative approach would be to ignore the missing components and to only use the present components to compute the HMM state likelihoods on the basis of marginal distributions. Experiments described in (Cooke et al., 1997) are related to multi-Gaussian HMM state modeling. The results in that study are in favor of the marginal approach which yields somewhat better results than the reconstruction approach.

However, the data reconstruction approach has several advantages. On the one hand, one can only use a limited number of Gaussians for the reconstruction part of the system. This allows to keep a compact system, without significant damage for the overall recognition system (at least during clean speech portions). On the other hand, it allows to obtain reconstructed vectors that (1) can be further processed (orthogonalisation, temporal filtering, ...), and (2) can be used as input to any classical ASR system. In our case, it will be an ASR system based on an ANN probability estimator.

4.3. Experiments

The purpose of this section is to test whether "accurate" noise estimates can contribute to an improvement of ASR system performances. The experiments will compare a non-adaptive noise estimator (noise is estimated from the first 100 ms

of each utterance and considered stationary throughout the utterance) to an adaptive noise level estimation method in the case of robust ASR setups using either spectral subtraction or the missing data paradigm. The adaptive noise estimator is the low-envelope method with 500 ms time segments and harmonic filtering. This method yielded the lowest overall MSE in the experiments reported in Section 3.

As a reference, we also tested J-RASTA features (Hermansky and Morgan, 1994), which are known to be very robust to additive noise. The J parameter was optimized for best overall recognition rate, leading to a value of 10^{-6}.

4.3.1. Spectral subtraction

As in (Berouti et al., 1979), β was set to 0.001. The α_0 parameter was optimized for best performance on speech data with additive white noise, leading to a value of 2. Following (Singh and Sridharan, 1998), spectral subtraction was applied to a filter-bank representation of the signal. We used a Bark scale analysis to obtain 15 channels covering the 0–4000 Hz frequency range, sampled at a frame rate of 10 ms. Note that this filter-bank analysis is only used to obtain parameters suitable for speech recognition purposes and that the noise level estimation still follows the setup described previously: we synthesize an estimation of the noise signal from 200-Hz frequency bands.

The complete feature vector was based on cube-root compressed (Hermansky, 1990) critical band energy values and on the first temporal derivative of the frame log-energy. Feed-forward neural networks were used to estimate the HMM-state posterior probabilities (Bourlard and Morgan, 1994) from 15 adjacent feature frames.

The TI-DIGITS (Leonard, 1984) corpus (American English digit sequences) was used for these experiments. Word models were build from 22 context independent phoneme models (including a silence model) composed of a single state. For each model, a minimum duration of half the mean duration of the phoneme was also used. Training was performed on a 1688 utterances set (2982 s of speech). Testing was performed on a 240 utterances subset of the database (1149 words). Two kinds of noise were added to the test utterances: the moving car noise from the MADRAS database ($N3$) and the factory noise from the NOISEX database ($N5$). These are the least stationary of the four real noises introduced in the previous sections. Speech and noise were added such that the mean energy of each utterance is SNR dB above the mean energy of the whole noise signal. Results for different SNR are plotted in Figs. 9 and 10.

4.3.2. Missing data

We used exactly the same spectral representation and the same hybrid HMM/ANN ASR system as for spectral subtraction. Tests were performed on the same data. The parametric model used for data reconstruction consisted of 16 multi-dimensional Gaussians with diagonal covariance matrices for modeling the whole data. This model is thus independent of the phonemes. Note that this configuration is very simple and was developed for the purpose of illustrating the application of noise level estimates. Better performance can be obtained with more complex configurations, such as full covariance matrices, phoneme-dependent models (as opposed to the 16 phoneme-independent Gaussian distributions used

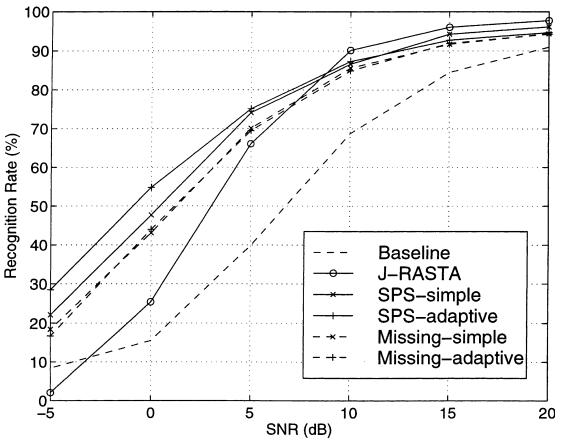

Fig. 9. Recognition rate (factory noise). 'Baseline' uses critical band features, 'SPS-simple' and 'missing-simple' use critical band features and noise spectrum estimation from the first 10 (100 ms) speech frames of each utterance, 'SPS-adaptive' and 'missing-adaptive' use adaptive noise spectrum estimation (low-energy envelope tracking, 500 ms speech segment and harmonic filtering).

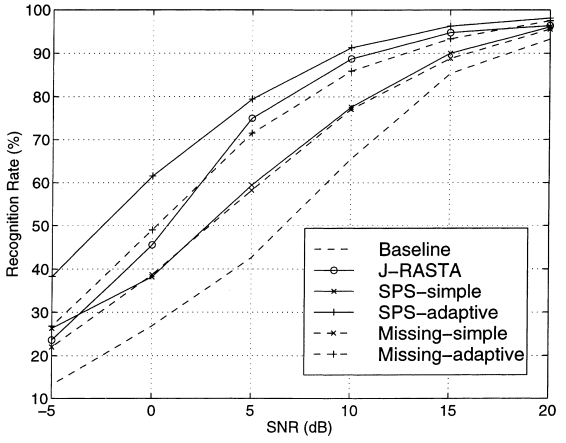

Fig. 10. Recognition Rate (MADRAS car noise). 'Baseline' uses critical band features, 'SPS-simple' and 'Missing-simple' use critical band features and noise spectrum estimation from the first 10 (100 ms) speech frames of each utterance, 'SPS-adaptive' and 'Missing-adaptive' use adaptive noise spectrum estimation (low-energy envelope tracking, 500 ms segment length and harmonic filtering).

in our experiments), bounded marginalisation (Cooke et al., 1999; Vizinho et al., 1999),... Results for different SNR are plotted in Figs. 9 and 10.

4.3.3. Results

We measured the word level accuracy as follows: recognition rate $= (N - S - I - D)/N$, where N is the number of words in the test set, S the number of substitution errors, I the number of insertions and D is the number of deletions. The significance regions at recognition rates of 90%, 70% and 50% are $\pm 1.7\%$, $\pm 2.6\%$ and $\pm 2.9\%$, respectively ($p = 0.95$).

Figs. 9 and 10 illustrate the benefits of combining adaptive noise spectrum estimation with robust ASR methods but also the limits of the noise spectrum estimation methods. Indeed, in the case of the NOISEX factory noise, even a fast adaptation of the noise level estimation cannot follow very fast or short changes of the noise conditions such as hammer blows or sudden siren wails. Therefore, the adaptive noise estimator does not signicantly improve non-adaptive methods and only the stationary part of the noise is estimated. In the case of the MADRAS car noise, the performance enhancement is noticeable as this noise presents amplitude modulations with rather low modulation frequencies (around 0.5 Hz). Note that for low SNR values (below 10 dB), the adaptive approach remarkably outperforms the J-RASTA method.

5. Conclusions

Robust ASR methods generally rely on good estimations of the noise statistics. In this paper, four noise level estimation procedures were compared. These methods can be used to estimate the noise level within narrow frequency bands. They avoid explicit speech pause detection and basically follow from the observation that relatively long speech segments hopefully contain non-speech portions that can be used to update the noise level estimates. It is shown in the paper that these methods perform well for stationary noise level estimation.

In the case of non-stationary noise, the alternation of speech and non-speech portions might not be quick enough to obtain reliable noise estimates. Assuming the signal energy in the valleys of a periodic sound spectrum is mainly due to background noise, we can use the periodic property of speech to update the noise level during voiced speech portions, without explicit detection of such portions. All of the compared algorithms can use this new method as an extension, yielding more accurate non-stationary noise level estimations.

In the same time, we tried to derive statistics on the optimal time segment duration, emphasizing the possibility to make this parameter frequency dependent. Indeed, the speech signal energy is mainly present around the first two formants. In lower and higher frequencies, *silent* portions (in the sense defined in this paper) are more frequent and longer thus allowing a faster adaptation of the noise level estimation in those frequency bands.

Finally, noise spectrum estimates were applied to spectral subtraction and missing data strategies in the framework of robust automatic speech recognition. These two pre-processing methods require an accurate estimation of the noise spectra. The study presented in this paper shows the

benefits we can draw from good noise spectrum estimation algorithms but also the limits of such methods. Indeed, in some noise conditions such as the factory noise used in this paper or the babble noise (this problem, known as *cocktail party*, has not been addressed in this paper), direct estimation methods (with no a priori knowledge on the noise properties) fail. Other noise robust ASR methods should be investigated to appropriately handle unfavorable cases. For instance, specific noise modeling, microphone arrays, ..., but this is, of course, beyond the scope of this paper.

Acknowledgements

This work is supported by the European Community long term research project RESPITE.

References

Berouti, M., Schwartz, R., Makhoul, J., 1979. Enhancement of speech corrupted by acoustic noise. In: Proc. ICASSP'79, April, pp. 208–211.

Boll, S.F., 1979. Suppression of acoustic noise in speech using spectral subtraction. IEEE ASSP 2 (27).

Bourlard, H., Morgan, N., 1994. Connectionist Speech Recognition – A Hybrid Approach. Kluwer Academic Publishers, Dordrecht, ISBN 0-7923-9396-1.

Bourlard, H., Dupont, S., Hermansky, H., Morgan, N., 1996. Towards sub-band-based speech recognition. In: Proceedings of European Signal Processing Conference, Trieste, Italy, September, pp. 1579–1582.

Cooke, M., Morris, A., Green, P., 1997. Missing data techniques for robust speech recognition. In: Proc. ICASSP'97, Munich, April.

Cooke, M., Green, P., Josifovski, L., Vizinho, A., 1999. Robust automatic speech recognition with missing and unreliable acoustic data. In: Research Memorandum CS-99-05, Department of Computer Science, University of Sheffield.

Dupont, S., 1998. Missing data reconstruction for robust automatic speech recognition in the framework of hybrid HMM/ANN systems. In: Proceedings of the International Conference on Spoken Language Processing, Sydney, Australia, December.

El-Maliki, M., Renevey, P., Drygajlo, A., 1998. Rehaussement par soustraction spectrale et compensation des param tres manquants pour la reconnaissance robuste du locuteur et de la parole. In: Proceedings XXII mes Journ es d'Etude sur la Parole, Martigny, Switzerland, pp. 409–412.

Gales, M.J.F., 1997. Nice model-based compensation schemes for robust speech recognition. In: Proceedings of ESCA/NATO Workshop on Robust Speech Recognition for Unknown Communication Channels, Pont–Mousson, France, April, pp. 55–64.

Hermansky, H., 1990. Perceptual linear predictive (PLP) analysis of speech. J. Acoustical Soc. Amer. 87 (4) 1738–1752, April.

Hermansky, H., Morgan, N., 1994. Rasta processing of speech. IEEE Trans. Speech Audio Processing 2 (4), 578–589.

Hirsch, H.G., Ehrlicher, C., 1995. Noise estimation techniques for robust speech recognition. In: Proc. ICASSP'95, pp. 153–156.

Korthauer, A., 1999. Robust estimation of the snr of noisy speech signals for the quality evaluation of speech databases. In: Proc. ROBUST'99 Workshop, Tampere, pp. 123–126.

Leonard, R.G., 1984. A database for speaker-independent digit recognition. In: Proc. ICASSP'84, pp. 111–114.

Martin, R., 1993. An efficient algorithm to estimate the instantaneous SNR of speech signals. In: Eurospeech'93, pp. 1093–1096.

McKinley, B.L., Whipple, G.H., 1997. Model based speech pause detection. In: Proc. ICASSP'97, Munich, pp. 1179–1182.

Mirghafori, N., Morgan, N., 1998. Combining connectionist multi-band and full-band probability streams for speech recognition of natural numbers. In: Proceedings of the International Conference on Spoken Language Processing, Sydney, Australia, December.

Price, P., Fisher, W.M., Bernstein, J., Pallet, D.S., 1988. The DARPA 1000-words resource management database for continuous speech recognition. In: Proc. ICASSP'88, April, pp. 651–654.

Sarikaya, R., Hansen, J.H.L., 1998. Robust speech activity detection in the presence of noise. In: Proceedings of the International Conference on Spoken Language Processing, Sydney, Australia, December.

Singh, L., Sridharan, S., 1998. Speech enhancement using critical band spectral subtraction. In: Proceedings of International Conference on Spoken Language Processing, Sydney, Australia, December.

Tibrewala, S., Hermansky, H., 1997. Sub-band-based recognition of noisy speech. In: Proc. ICASSP'97, Munich, pp. 1255–1258.

ULg, 1998. ULg – acoustics laboratory – the MADRAS project. http://www.montefiore.ulg.ac.be/services/acous/Cedia/madrasfr.html, August.

Van Compernolle, D., 1989. Noise adaptation in a hidden markov model speech recognition system. Comput. Speech Language 3 (2), 151–168.

Varga, A.P., Steeneken, H.J.M., Tomlinson, M., Jones, D., 1992. The NOISEX-92 study on the effect of additive noise on automatic speech recognition. Technical Report, Speech Research Unit, Defense Research Agency, Malvern, UK.

Vizinho, A., Green, P., Cooke, M., Josifovski, L., 1999. Missing data theory, spectral subtraction and SNR estimation for robust ASR: An integrated study. In: Proc. EUROSPEECH'99, Budapest, pp. 2407–2410.