Robust speech recognition to non-stationary and unpredictable noise based on model-driven approaches

Christophe Cerisara, Irina Illina

Robust speech recognition to non-stationary and unpredictable noise based on model-driven approaches

Christophe Cerisara, Irina Illina

CNRS/LORIA - UMR 7503
54506 Vandoeuvre-les-Nancy, FRANCE
{cerisara, illina}@loria.fr

Abstract

Automatic speech recognition works quite well in clean conditions, and several algorithms have already been proposed to deal with stationary noise. The next challenge probably consists to compensate for non-stationary noise as well. This work studies this problem by proposing and comparing two adaptations of the Parallel Model Combination (PMC) algorithm for non-stationary noise. A third method, derived from the missing data framework, is further proposed and compared to the two previous ones. In musical noise, experimental results show an important improvement of the recognition accuracy for one PMC-derived algorithm, compared to the non adapted system. The missing-data algorithm also performs quite well, despite its simplicity and the strong assumptions he is using.

1. Introduction

Traditional PMC algorithm assumes that the noise model that represents the noise present in the test sentences is known a priori. This assumption presents the following drawbacks:

- It is not often easy to know in realistic situations which kinds of noise may actually corrupt the speech signal. And even when we know for example that a passing car may produce some noise, it is not obvious whether our stored car noise models will accurately represent the precise car noise that is passing at that given time.
- Even when the acoustic models are insensitive to amplitude mismatch (for example cepstral models without c_0), the ratio between the spectral power of the speech and noise must be known, as the adaptation is realized in the power spectrum. It is not obvious to compute this ratio when the noise models have been trained a priori, for example on a noise database.
- It is assumed that the speech and noise power spectra are additive, but this is only an approximation of the true combination scheme, and it is known that this approximation might introduce some error [8].

To address the latter issue, other speech and noise combination equations can be used, for example by consider a masking rather than an additive scheme, as it is done in missing data recognition [1]. But this is once again another approximation.

To address the two former issues, the noise models are usually not trained a priori on an independent database, but are rather estimated on the test sentence itself, for example during the silence segments of the signal. Therefore, the ratio of the noise and "silence" energies is implicitly known, and the noise model precisely represents the noise that actually contaminates the test sentence. However, this solution (i) requires a good segmentation of the incoming signal into silence and noise segments and (ii) limits the use of PMC to "quasi-stationary" noises, as the noise is assumed to be constant between two silence segments.

We propose in this work to test and compare the two solutions to train the noise models, and we propose a method to use PMC without assuming that the noise is constant between two silence segments. Section 2 proposes a method to use a priori trained noise models, while section 3 adapts PMC to non-constant noises. Section 4 briefly presents the noise tracking algorithm that is used in both algorithms. Section 5 describes a simple method derived from the missing-data framework, which is known to be especially robust to non-stationary noises. This algorithm is far from the most advanced methods in the very active field of missing data recognition, but it is mainly used for comparison purposes. Section 6 presents our experimental results and section 7 concludes the paper.

2. Adaptation Based on a Priori Noise Models

We propose in this section a method to use a priori trained noise models within the PMC framework. The main issue concerns how to estimate the energy ratio between the speech power and the noise power.

It is possible to handle this problem by considering a multiplicative factor for the noise model. Let S and N be respectively a speech and noise Gaussians in the power spectral domain. Let O be the power spectral observation vector with which S and N are aligned. The idea is to combine the speech and noise models in the power spectral domain with the following equation:

$$Z = S + \alpha N \tag{1}$$

Given a frame-state alignment, it is possible to compute an optimal α for each frame that minimizes, for example, the mean square error criterion:

$$\hat{\alpha} = \arg\min_{\alpha} \sum_i (S_i + \alpha N_i - O_i)^2 \tag{2}$$

where the index i represents a given scalar coefficient of the vector.

However, our preliminary experiments have shown that this procedure may give too much flexibility in the use of the noise model. We thus propose to compute a single α at each beginning of sentence, using the following method:

- A "running estimate" of the magnitude of the instantaneous noise is computed for every frame of the sentence, using the algorithm described in section 4. Let $|N(t)|$ be the value of this running estimate at time t.

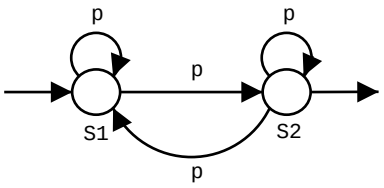


Figure 1: Noise model

- α represents the relative energy of the target noise with respect to the training noise and is computed by:

$$\alpha = \left(\frac{\max_t(|N(t)|)}{|N_0|} \right)^2 \quad (3)$$

where $|N_0|$ is the magnitude of the training noise.

In our experiments, we use a Gaussian Mixture Model (GMM) to model the noise. This GMM is trained on a noise database that contains the same kinds of noise that the noise which corrupts the test signal, but of course not the same noise files. Also, we want to deal with unpredictable noise, in the sense that the noise may or may not occur in the sentence. We have introduced this constraint into our algorithm by creating a two-states HMM for the noise model, as the one represented in figure 1.

The state $S1$ represents silence and should align with the "unnoisy" segments of the signal, while the state $S2$ is actually the noise GMM that has been trained on the noise database. The transition probabilities are all equal to $p = 0.5$. The emitting probability of state $S1$ is a constant power spectral vector equal to zero. This means that, when combined with a speech power spectral vector, it does not alter the speech model at all. Due to the particular topology of this noise model, the combination of this noise model with a classical left-to-right speech HMM can be realized by simply duplicating every Viterbi path: one path will align with state $S1$ and its clone with state $S2$. We further use the *max* operation to compute the emission probability of a frame aligned with state $S2$ instead of the classical weighted sum for that mixture. This comes from the assumption that each mixture in the noise GMM represents a different kinds of noise of the training database. Thus, during testing, only one of these possible noise may occur, and most probably not all of them simultaneously. The cost of the search procedure is thus equal to $N_m + 1$ times the cost of a classical Viterbi algorithm, where N_m is the number of mixtures of the original noise GMM.

Summary of the approach:

In this approach, the target noise is not directly estimated, but is assumed to belong to a noise database which contains every possible noise that may occur in a given environment. Each noise *may* be combined with the speech model aligned with one frame. Every possible combination between this frame and a noise Gaussian is treated in parallel by the decoding algorithm. Furthermore, the energy of the noise model is adapted to the current sentence using the α factor described above.

Based on these considerations, we have decided to name this algorithm *static OPMC* , for *Optional* PMC with a *static* noise model, by contrast with section 3 that uses a dynamic noise model.

3. Adaptation Without a Priori Noise Models

Because of the problems exposed in section 1 concerning the use of a priori noise models, we adapt the previous algorithm by estimating the noise on the test sentence itself, which is closer from what is usually done in model adaptation.

The characteristics of our method are the following:

- The noise model is combined *optionally* with the speech models, as it is explained in section 2. The noise model represents potential bursts of sounds, rather than the continuous background sound to which acoustic models are traditionally adapted.
- We try to obtain precise representations of every noise source that occur during the sentence. To achieve this, we have used the noise magnitude estimator, described in the previous section. Then, all the detected noise frames are classified into M Gaussian mixtures.
- The evolution of the noise in time is not important here, as if a given noise actually changes between t and $t + 1$, then the system will consider that two different noises occur at t and $t + 1$.

The basic principle of this algorithm can be summarized into the following steps:

1. All the noise frames are extracted from the incoming signal and a GMM noise model is estimated on these frames;
2. The OPMC algorithm described in the previous section is then applied with this estimated noise model.

The OPMC algorithm that is used here is very similar to what is described in section 2, except that no α factor is needed. Indeed, this factor was used to make the energy of the a priori noise model match the energy of the target noise, but this is not required any more here, as the noise is directly estimated from the test signal. This algorithm is called *dynamic OPMC*.

4. Noise tracking algorithm

To estimate the noise directly on the test sentence, a noise tracking algorithm is required to identify the regions dominated by noise. We use in this work an algorithm derived from [6]. The algorithm basically segregates speech and noise segments based on an energy criterion. Our main modification compared to [6] consists to add a second pass to the algorithm to fix some errors that it is doing at the boundaries of speech segments: a few frames corresponding to the beginning and to the end of the speech segments are often affected to noise, whereas they should be affected to speech.

An example of the segmentation of one sentence corrupted by musical noise is shown in figure 2.

Of course, other noise tracking algorithms than the one presented here can be used (and actually should be used when the conditions differ) with both previous adaptation algorithms, like for example by using the union probabilistic model [9] or the FSVA algorithm [10].

5. Model-Driven Missing Data Recognition System

We consider in this section a missing-data recognition algorithm, which makes use of masks to "hide" the parts of the time-frequency plane that are considered as noisy. We do not

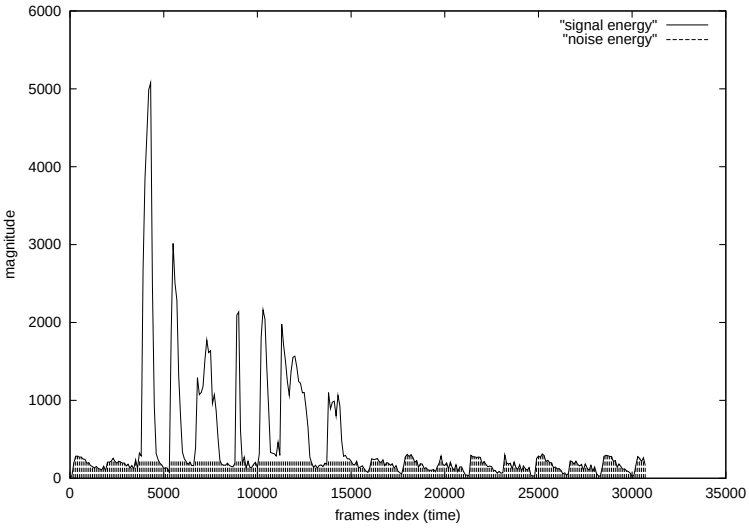
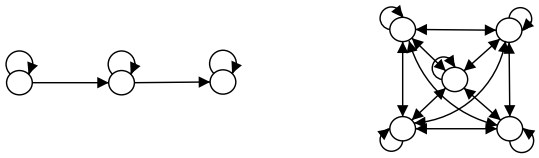


Figure 2: Example of the segmentation of the sentence “trente neuf mille sept cent quarante sept” into speech and noise segments. The curve represents the local energy while the dashed region represents the estimated noise energy.

use any a priori noise model in this section. This algorithm is described next from a model-combination point of view.



(a) speech HMM (b) mask HMM

Figure 3: Models used in the model-driven missing data approach

The speech units are modeled like usual by left-to-right HMMs, whereas the boolean masks are generated by an ergodic HMM like the one represented in figure 3(b). This model defines every possible mask that can be applied at any instant. Each state of this model generates one such mask.

The recognition algorithm simply combines the speech and mask models, and then realizes a traditional Viterbi decoding to maximize the likelihood of the observations. The combination of a mask vector and a speech Gaussian is realized in the missing-data framework, by considering only the unmasked parts of the power spectrum, as it is described in [1].

Usually, the masks are built using “bottom-up” signal processing techniques, for example the computation of local SNRs. A better approach consists to combine bottom-up procedures with top-down inference, as it is realized in the multi-source decoder [2]. We rather test here a purely top-down approach, where the decoding process only chooses the best masks by maximizing the recognition score.

The other characteristics of our method are:

- We use MFCC parameters, but the masks are applied in the spectral domain, in a similar way as it is realized in [11]. This choice creates a lot of theoretical and practical problems that are discussed in a companion paper.
- We use 4 frequency bands, defined by splitting the Mel-scale filterbanks into 4 groups of same size. We a priori define only 5 possible masks: the full-band and 4 masks in which one different sub-band is masked. These masks have been chosen to control the complexity of the algorithm (which is equal to 5 times the complexity of Viterbi with a priori masks) and to constraint the system not to lose too much acoustic information because of the mask.
- We use a boolean mask with “hard” decision, that is a spectral coefficient is either considered as is or not considered at all. Soft decisions have proven to be better [1], but it is very difficult to use them with MFCC coefficients.
- To compare the scores returned by each mask, we apply the a posteriori normalization technique, as it is suggested in [2], rather than a scaling factor.

In this work, the masks are generated using a very simple model, but in future work, such models could be trained on some database, just like speech models are. A first extension would be to train the transition probabilities of the mask model, which would prevent the masks from changing too often. However, we have not investigated these directions in this study.

6. Experiments

The task consists to recognize unconstrained sequences of French numbers. 27 words models are used to represent the French numbers. The SPEECHDAT (telephone) database has been used for training and testing. A background music has been added to the test corpus at different SNRs. This type of noise has been chosen because it is non-stationary and unpredictable, in the sense that the system does not know when drum sounds occur, for example (successive time frames are assumed uncorrelated). Speech models are 13-emitting states HMM with 8 Gaussians per state. Acoustic vectors have $13+\Delta+\Delta\Delta$ MFCC coefficients. For static OPMC, the a priori noise model is built on another music, from a different artist than the one chosen for testing. For dynamic OPMC, the noise model is built by clustering all the frames that are given by the noise tracking algorithm with the LBG algorithm into 16 classes ($M = 16$).

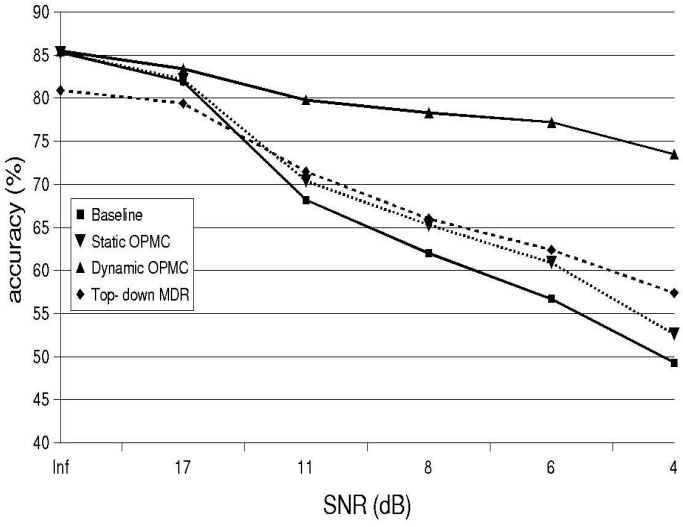


Figure 4: Experimental results in musical noise

Every method proposed here presents better results than the baseline (non adapted) system, but the best algorithm is clearly dynamic OPMC. The MDR system performs also quite well at low SNR, despite its simplicity, and the fact that it does not use any noise model.

Remarks on the Dynamic OPMC Algorithm:

Although this method gives the best results in these tests, other conditions may impair its effectiveness:

- *Important stationary noise*: when the noise is always present and has a relatively high level compared to speech, then the noise tracking algorithm may not manage to differentiate noisy and speech segments, and may classify everything as noise. This issue may be addressed by combining the scheme proposed here with adaptation methods robust to stationary noise.
- *Noise which occur only in speech fragments*: when the noise occurs only in speech fragments, then the noise tracking algorithm might not detect and model it.

7. Conclusions

The main contributions of this paper are the following:

- Proposition of two algorithms to adapt PMC to non-stationary and unpredictable noise;
- Derivation of a third model-driven MDR algorithm that uses MFCC features and a posteriori probability normalization;
- Evaluation and comparison of these three algorithms on a telephone database with added musical noise.

The three algorithms proposed in this work are model-driven, which means that the decomposition of the acoustic signal into its sub-components (one per source) is achieved by the models of speech and noise. Although the dynamic OPMC algorithm clearly takes the advantage in experimental results, the missing data recognition method performs reasonably well compared to its simplicity. More advanced missing data algorithms can be used, but our constraints imposed us to use MFCC models, and there are still several important issues to solve before applying missing data techniques with such models. However, we do not address these issues here, but let these points for a companion paper. Also, the low scores obtained with static OPMC suggest that the issues mentioned in the introduction relating to the use of a priori noise models still needs a lot of working to be solved.

Finally, we would like to point out that the usability of the two first algorithms strongly depend on the noise tracking algorithm that is used, and other methods should probably be considered in different conditions.

8. Acknowledgements

This work was supported by the IST 2000-30026 OZONE EC project:
(http://www.extra.research.philips.com/euprojects/ozone/).

9. References

[1] M. Cooke, P. Green, L. Josifovski, and A. Vizinho, "Robust Automatic Speech Recognition with Missing and Unreliable Acoustic Data," *Speech Communication*, vol. 34, no. 3, June 2001.

[2] J. Barker, M. Cooke, and D. Ellis, "Decoding Speech in the Presence of Other Sound Sources," in *ICSLP'00*, Beijing, China, 2000.

[3] D. Ellis, *Prediction-driven Computational Auditory Scene Analysis*, Ph.D. thesis, EECS dept., M.I.T., 1996.

[4] M.J.F. Gales, *Model-Based Techniques For Noise Robust Speech Recognition*, Ph.D. thesis, Gonville and Caius College, September 1995.

[5] C. Cerisara and D. Fohr, "Multi-Band Automatic Speech Recognition," *Computer Speech and Language*, vol. 15, no. 2, pp. 151–174, Apr. 2001.

[6] H.-G. Kim and D. Ruwisch, "Speech Enhancement in Non-Stationary Noise Environments," in *ICSLP'02*, Denver, USA, September 2002.

[7] C. Cerisara, J.-C. Junqua, and L. Rigazio, "Dynamic Estimation of a Noise over Estimation Factor for Jacobian-Based Adaptation," in *ICASSP 2002*, Orlando, Floride, May 2002.

[8] J. Droppo, A. Acero, and L. Deng, "A Nonlinear Observation Model for Removing Noise from Corrupted Speech Log Mel-Spectral Energies," in *ICSLP 2002*, Denver, Colorado, pp. 1569–1572, September 2002.

[9] J. Ming and F.J. Smith, "Union: a Model for Partial Temporal Corruption of Speech," *Computer Speech and Language*, vol. 15, pp. 217–231, 2001.

[10] M. Siu and Y.-C. Chan, "Robust Speech Recognition Against Short-Time Noise," in *ICSLP'02*, Denver, USA, September 2002.

[11] J. Häkkinen and H. Haverinen, "On the Use of Missing Feature Theory with Cepstral Features," in *CRAC Workshop*, Aalborg, Denmark, September 2001.