

Robust speech recognition to non-stationary 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 consists to work with non-stationary noise. This paper studies this problem. We propose three algorithms to non-stationary noise adaptation : Static and Dynamic Optional Parallel Model Combination (OPMC) and one algorithm derived from the Missing Data framework. The combination of speech and noise is expressed in the spectral domain and different ways to estimate the non-stationary noise model are studied. The proposed algorithms are tested on a telephone database with added background music at different SNRs. The best result is obtained using dynamic OPMC.

1. Introduction

Automatic speech recognition systems give very good results when the testing conditions match the training context, including environment noise, microphones and speakers. Several adaptation techniques such as PMC [1, 2] and MLLR have been proposed to deal with noise mismatch. However, these techniques usually require the noise to be quasi-stationary and the noise model to be known or estimated using the adaptation speech. If the environment mismatch between training and testing is non stationary, recently proposed Missing Data approach (MDA) [3] can be used. It first detects the spectro-temporal regions that are not contaminated by noise and then adapts the recognition algorithms to make use of this information. The regions contaminated by noise are considered as missing.

We study in this work both approaches to handle noisy speech recognition: Section 2 presents the baseline PMC advantages and drawbacks that motivate our contribution; Section 3 proposes an “extension” of PMC to handle non-stationary noise, while section 4 proposes one possible implementation of the missing data approach. Section 5 compares these methods and section 6 concludes the paper.

2. Classical PMC

2.1. On noise estimation

In the traditional description of PMC, the noise model is assumed to be known *a priori*. This assumption presents the following drawbacks:

- In realistic situations, it is not easy to know exactly which kinds of noise might happen and how they may corrupt the speech signal.

- As the adaptation of the speech model to the noise is realized in the power spectrum domain, it is required to know the ratio between the powers of the speech and of noise. This ratio is usually very difficult to obtain.
- It is assumed that the speech and noise power spectra are additive, but this is only an approximation of the reality [4].

Another possibility to estimate the noise model is to use the test sentence itself, for example during the silence segments of the signal. In that case, the ratio of the noise and “silence” powers is implicitly known, and the noise model represents the noise that actually contaminates the test sentence. But this solution requires a good segmentation of the incoming signal into silence and noise segments.

We will study and compare in the rest of the paper both approaches to estimate the noise models for the algorithm presented in section 3.

2.2. On noise stationarity

When the speech and noise models have the classical HMM topology, PMC can only handle stationary or quasi-stationary noise. A quasi-stationary noise is for example a stationary noise between two successive silence segments (eventually with a short duration), or a noise that presents a repetitive structure that can easily be modeled by a HMM (such as a machine gun firing). But “unpredictable” noise like a door slamming can hardly be considered within this classical framework. We describe in the next section a particular case of PMC where the noise model is adapted in order to handle such non-stationary noise.

3. Optional PMC

3.1. Topology of the noise model

In this section, an adaptation of the classical PMC framework is described to handle non-stationary noise. The basic idea consists to create a particular noise model that takes into account different kinds of noise; During recognition, each frame may or may not be contaminated by noise. The topology of this noise model is presented in figure 1.

In this figure, the state $S1$ represents silence and should align with the “clean” parts of the signal, while the state $S2$ is a noise Gaussian Mixture Model (GMM) that models every possible noise that may occur during recognition. We will see in the next sections how this GMM can be trained. We assume that the transition probabilities p are all equal to 0.5. The emitting GMM in 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

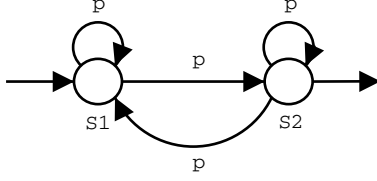


Figure 1: Noise model

its particular topology, 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 $S2$.

3.2. Computation of the emission likelihoods

We assume for the noise model described in the previous section that each Gaussian in the noise GMM models a different noise. We further assume that during testing, only one of these possible noise may occur at a given time. Finally, we use the \max operation instead of the classical weighted sum to compute the emission probability of a frame aligned with noise state $S2$.

Using this particular noise model, it is easy to see that adaptation of the speech models is realized *optionally*. Based on these considerations, we called this algorithm *Optional PMC* (OPMC). The noise model may represent bursts of sounds as well as continuous sounds. Also, when the noise continuously changes between instants t and $t + 1$, then the system will consider that two different noises occur at t and $t + 1$.

Finally, the cost of the search procedure during the testing is equal to $N_M + 1$ times the cost of a classical Viterbi algorithm, where N_M is the number of mixtures of the noise GMM in the state $S2$.

Two versions of OPMC are proposed next: *Static OPMC* uses *a priori* trained noise model, whereas *dynamic OPMC* estimates the noise model on the test sentence.

3.3. Static Optional PMC

In this section, a method to use *a priori* trained noise models for OPMC is proposed. One important issue concerns how to estimate the energy ratio between the speech power and the noise power. We consider next a gain factor for the noise model.

Let S and N be respectively a speech and noise models in the power spectral domain. We assume that S and N are trained *a priori*. Let O be the power spectral observation vector with which S and N are aligned. The speech and noise models are combined in the power spectral domain with the following equation:

$$Z = S + \alpha N \quad (1)$$

where α is the unknown gain factor, which represents the mismatch between the powers of the train and test noise.

We propose to compute α at the beginning of each test sentence, using the following method:

- A “running estimate” of the magnitude of the instantaneous noise is computed for every frame of the sentence, using a noise tracking algorithm described in section 3.5. Let $|N(t)|$ be the value of this running estimate at time t .
- α represents the ratio of the power of the target noise

with the training noise, and is computed by:

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

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

We call this version of OPMC *static OPMC* because it uses static noise model estimated *a priori*. 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.

3.4. Dynamic OPMC

Another solution to train a noise model consists to estimate it on the test sentence itself: This is what is usually done in model adaptation. Our implementation of this algorithm can be summarized into the following steps:

1. All the noise frames are extracted from the incoming signal using the noise tracking algorithm presented in section 3.5;
2. A GMM noise model is trained on these frames;
3. The OPMC algorithm is then applied with this estimated noise model on the current sentence;
4. This procedure is repeated for the next sentence with a new GMM.

The OPMC algorithm that is used here is very similar to what is described in section 3.3, 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*.

3.5. 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 [5]. It is assumed that the noise is not correlated with the clean speech. Thus, the noise component is additive in the magnitude spectrum of the signal. The algorithm basically segregates speech and noise segments based on the ratio of the noisy speech and its minimum. The algorithm has a very low computational requirements.

We modified this algorithm by adding a second pass to fix some errors at the boundaries of speech segments: We indeed observed that 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. We fixed these mistakes by extending in the second pass the segments affected to speech up to the very beginning of the increase and decrease of the magnitude of the noisy signal.

An example of the segmentation into noise and speech is shown in figure 2. The speech sentence is corrupted by musical noise.

Of course, other noise tracking algorithms, such as the union probabilistic model [6] or the FSVA algorithm [7], can be used for OPMC.

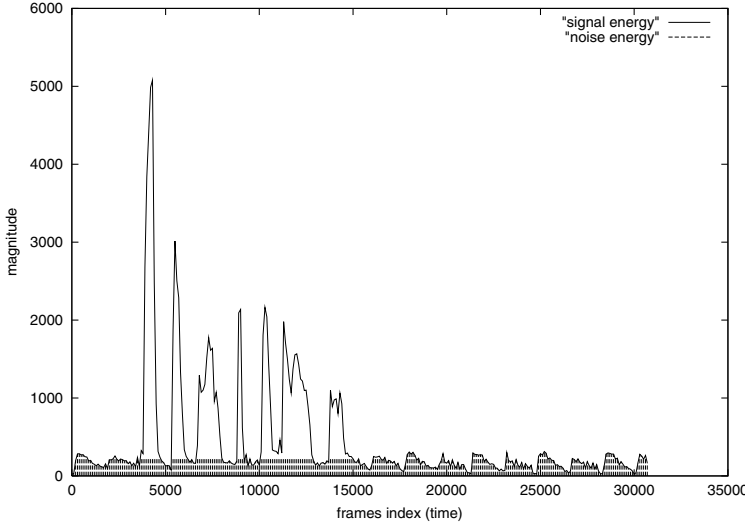


Figure 2: Example of the segmentation of the sentence “trente neuf mille sept cent quarante sept” into speech and noise segments. The curve represents the local energy while the dashed region represents the estimated noise energy.

4. Missing Data Approach

In this section, a missing-data recognition algorithm, which makes use of masks to “hide” the noisy parts of the time-frequency plane is described. We propose this approach because MDA has received much attention thanks to its faculty to deal with non-stationary noise, and to compare with the algorithms proposed in previous sections. Many different recognition algorithms exist in the field of Missing Data recognition, and we propose next a new but simple one that does not use any preprocessing of the signal.

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 [8]. Our algorithm uses a purely top-down approach, where the masks are generated by a mask model represented in figure 3. The decoding process only chooses the best masks by maximizing the recognition score.

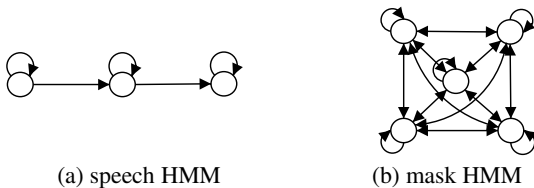


Figure 3: Models used in the Missing Data approach

The recognition algorithm combines the speech and mask models, and realizes a traditional Viterbi decoding to maximize the likelihood of the observations. The combination of a mask vector and a speech Gaussian is realized within the MDA framework, by considering only the unmasked parts of the power spectrum, as it is described in [3].

The 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 [9].

- 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 sub-band is masked.
- The speech units are modeled by left-to-right HMMs (figure 3(a)), whereas the boolean masks are generated by an ergodic HMM (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.
- We use a boolean mask with “hard” decision : a spectral coefficient is either considered as is or not considered at all. Soft decisions have proven to be better [3], 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 [8], rather than a scaling factor.

5. Experiments

The SPEECHDAT (telephone) database has been used for training and testing. The task consists to recognize unconstrained sequences of French numbers pronounced by native speakers. A background music has been added to the test corpus at different SNRs. Acoustic vectors have $13+\Delta+\Delta\Delta$ MFCC coefficients. 27 words models are used to represent the French numbers. Each word model is a 13-states HMM with 8 Gaussians per state.

For static OPMC, the *a priori* noise model is trained on music files of the same type but different than the one chosen for testing. For dynamic OPMC, the noise model is trained using the frames given by the noise tracking algorithm on the test sentence. These frames are clustered with the LBG algorithm into 16 classes ($M = 16$).

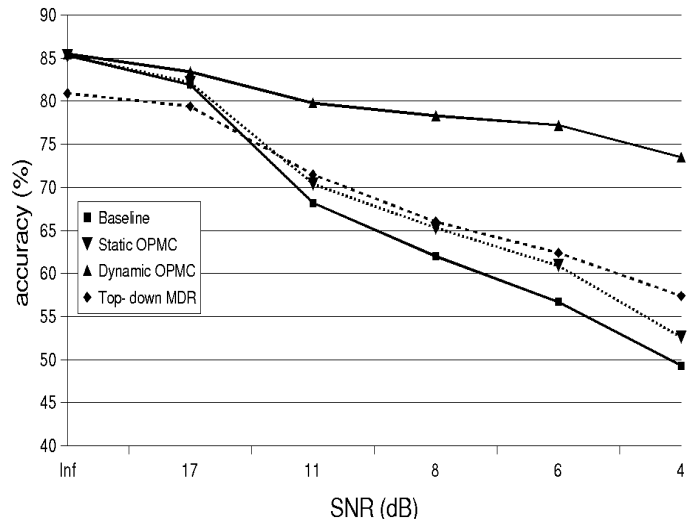


Figure 4: Experimental results in musical noise

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

6. Conclusions and perspectives

The main contributions of this paper are the following:

- Proposition of two algorithms to adapt PMC to non-stationary noise;
- Derivation of a 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 background 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/masks. 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.

The following points still need to be addressed in future work:

- *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.

The usability of OPMC algorithms strongly depends on the noise tracking algorithm that is used, and other methods should probably be considered in different conditions.

7. Acknowledgements

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

8. References

- [1] M.J.F. Gales, *Model-Based Techniques For Noise Robust Speech Recognition*, Ph.D. thesis, Gonville and Caius College, 1995.
- [2] M.J.F. Gales, *Predictive Model-Based Compensation Schemes for Robust Speech Recognition*, *Speech Communication*, vol. 25, no.1, pp 49-74, 1998.
- [3] 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, 2001.
- [4] J. Droppo, A. Acero, and L. Deng, "A Nonlinear Observation Model for Removing Noise from Corrupted Speech Log Mel-Spectral Energies," in *ICSLP 2002*, pp. 1569–1572, 2002.
- [5] H.-G. Kim and D. Ruwisch, "Speech Enhancement in Non-Stationary Noise Environments," in *ICSLP'02*, pp. 1829-1832, 2002.
- [6] 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.
- [7] M. Siu and Y.-C. Chan, "Robust Speech Recognition Against Short-Time Noise," in *ICSLP'02*, pp. 2373-2376, 2002.
- [8] J. Barker, M. Cooke, and D. Ellis, "Decoding Speech in the Presence of Other Sound Sources," in *ICSLP'00*, 2000.
- [9] J. Häkkinen and H. Haverinen, "On the Use of Missing Feature Theory with Cepstral Features," in *CRAC Workshop*, 2001.
- [10] D. Ellis, *Prediction-driven Computational Auditory Scene Analysis*, Ph.D. thesis, EECS dept., MIT, 1996.
- [11] C. Cerisara and D. Fohr, "Multi-Band Automatic Speech Recognition," *Computer Speech and Language*, vol. 15, no. 2, pp. 151–174, 2001.
- [12] C. Cerisara, J.-C. Junqua, and L. Rigazio, "Dynamic Estimation of a Noise over Estimation Factor for Jacobian-Based Adaptation," in *ICASSP 2002*, pp. 201-204, 2002.