

MASK ESTIMATION FOR MISSING DATA RECOGNITION USING BACKGROUND NOISE SNIFFING

Sébastien Demange, Christophe Cerisara and Jean-Paul Haton

LORIA - UMR 7503

54500 Vandoeuvre-les-Nancy - FRANCE

demangs@loria.fr, cerisara@loria.fr, jean-paul.haton@loria.fr

ABSTRACT

This paper addresses the problem of spectrographic mask estimation in the context of missing data recognition. At the difference of other denoising methods, missing data recognition does not match the whole spectrum with the acoustic models, but rather considers that some time-frequency pixels are missing, i.e. corrupted by noise. Correctly estimating these “masks” is very important for missing data recognizers. We propose a new approach that exploits some *a priori* knowledge about these masks in typical noisy environments to address this difficult challenge. The proposed mask is then obtained by combining these noise dependent masks. The combination is led by an environmental “sniffing” module that estimates the probability of being in each typical noisy condition. This missing data mask estimation procedure has been integrated in a complete missing data recognizer using bounded marginalization. Our approach is evaluated on the Aurora2 database.

1. INTRODUCTION

Robustness of automatic speech recognition to noise is a major challenge for nowadays state-of-the-art systems. While a number of efficient methods have been proposed to deal with quasi-stationary or slowly varying noise, very few techniques can handle non-stationary and highly variable noise. Missing Feature Theory (MFT) is such an approach, which assumes that the noise masks the speech signal in some localized spectrographic regions.

A typical MFT speech recognizer is composed of two stages:

The first step produces a mask that identifies the reliable spectrographic features (dominated by speech), and the corrupted features (dominated by noise). This mask is defined for every frame and every spectral coefficient: it is thus composed of $T \times N$ boolean values, where T is the total number of frames and N is the number of frequency bands at the output of the front-end. *Soft masks* have also been proposed [1] to encode the probability that each feature is corrupted by noise instead of the hard reliable/corrupted decision.

The second step recognizes the speech by taking into account the mask. Two different approaches can be used The first one called *marginalization*, marginalizes the observation likelihood of the corrupted (hidden) observations. Several variants of marginalization have been proposed; a typical one is the following:

$$P(X|\Theta) = P(X_r|\Theta) \cdot \int P(X_m|\Theta) dX_m \quad (1)$$

where X_r and X_m are the reliable and missing coefficients of the feature vector X , and Θ the acoustic models. The second one called *data imputation* estimates the contribution of speech in the masked observations.

A number of publications have shown that MFT significantly improves the performance of speech recognizers under noisy conditions when the masks are known *a priori* [2]. However, this *a priori* information is not available in real conditions. Hence, a major challenge is to estimate accurately the missing data mask. This paper is focused on this issue.

We propose a two step method. First, several missing data masks are computed from mask estimators trained on different *a priori* known environments. These environment dependent masks are then combined to give the final mask. The combination is led by an environment “sniffing” module that gives the probability that the unknown environment is one of the *a priori* environments.

The organization of the paper is as follows. In section 2, we present an overview of some state-of-the-art mask estimation techniques. In section 3, we expose our approach. Section 4 presents the experimental setup, and section 5 summarizes results. In section 6, some limitations of the proposed method are discussed and possible improvements are suggested. Finally, conclusions and future work are given in section 7.

2. RELATED WORKS

The litterature proposes a wide range of methods that estimate the masks. While some of them are motivated by psycho-acoustical considerations, others are derived from signal processing algorithms. The most important of these methods are summarized next.

Drygajlo and El-Maliki [3] proposed to exploit spectral subtraction to estimate the missing data masks. The main drawback of this approach is that spectral subtraction works well for stationary noise, but fails to capture accurately non-stationary noise.

It was also shown that the local signal-to-noise ratio (SNR) is a good criterion to classify spectrographic features as reliable or missing. However, it is difficult to correctly estimate the local SNR since the clean speech signal is usually unknown. Nevertheless, techniques based on thresholding of an estimate of local SNR (*a posteriori* SNR) constituted the first attempts to automatically infer the missing data masks [4]. In this thesis Renevey showed that spectral subtraction and SNR based detections are equivalent.

Another classical criterion used to estimate the masks is the harmonicity measure, which is combined in [5] and [6] with the local SNR.

In [7], a neural oscillator network is also proposed to compute the missing data mask.

Roweis introduced an approach based on the factorial-max vector quantization model in [8]. A set of noise and speech codebooks were combined using the masking approximation, which stated that only one codebook dominates in a given frequency band and at a given frame. The resulting masks were thus obtained by finding the sequence of noise and speech codebooks that maximized the likelihood of the noisy sentence.

A top-down procedure is presented in [9] to automatically select the most appropriate masks within a set of potential candidates derived with the previous techniques.

Seltzer proposed a Bayesian classifier to label spectrographic features [10], which does not assume any prior knowledge about the noise. While Seltzer used speech corrupted by white noise to train the classifier, Kim *et al.* proposed in [11] a new training method based on subbands of colored noise.

3. COMBINATION OF ENVIRONMENT DEPENDENT MASKS

The proposed approach relies on bayesian classification of spectrographic features into two classes: reliable and missing. We propose to train several bayesian mask estimators and to combine their classification decision. Each mask estimator is trained on a particular corrupting noise at a given SNR. An environmental “sniffing” module is used to combine these specialized mask models.

We consider a finite set $E = \{e_1, \dots, e_K\}$ of representative environments that are known *a priori*. Each e_k is characterized by its nature (car noise, street noise, ...) and by its SNR. Inspired by [12], we propose to compute the probability $p(e_k|y)$ that e_k is the background noise of frame y .

The basic principle of our approach is illustrated by the graphical network of figure 1.

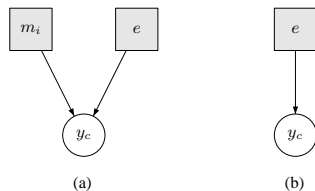


Fig. 1. Graphical representation of the dependencies in (a) our mask models and (b) our environment models. Square nodes are discrete, and greyed ones are hidden.

On this figure, m_i is a boolean variable that represents the mask for the i^{th} coefficient of the spectral observation y , y_c is the cepstral parametrization of y , and $e \in E$ is a discrete variable that represents the test environment. Two Gaussian mixture models (GMM) are considered, one for the masked and reliable coefficients given some noisy environments, and some to detect the environment. Both GMMs are defined in the cepstral domain.

The overall method can be decomposed into two steps:

Step 1: Noise dependent mask estimation

In this step, we assume that the background environment e_k is known. The probability that the i^{th} spectral coefficient is masked is then

computed:

$$p(m_i|y_c, e_k) = \frac{p(y_c|m_i, e_k) \cdot p(m_i|e_k)}{p(y_c|m_i, e_k) \cdot p(m_i|e_k) + p(y_c|\overline{m}_i, e_k) \cdot p(\overline{m}_i|e_k)} \quad (2)$$

where $p(m_i|e_k)$ and $p(\overline{m}_i|e_k)$ respectively represent the *a priori* probability that the i^{th} feature of y is missing and reliable in the environment e_k . $p(y_c|m_i, e_k)$ and $p(y_c|\overline{m}_i, e_k)$ are modeled by two GMMs.

Step 2: Combination with environment sniffing

In this stage, the probability that y_c is corrupted by an *a-priori* environment e_k is given by:

$$p(e_k|y_c) = \frac{p(y_c|e_k) \cdot p(e_k)}{\sum_{j=1}^{card(E)} p(y_c|e_j) \cdot p(e_j)} \quad (3)$$

We assume that all the environments have the same *a priori* probability: $p(e_i) = p(e_j) \forall (i, j)$. The likelihood $p(y_c|e_k)$ is given by a GMM. We propose to merge masks computed in the first step according to these environment probabilities. The combination scheme is a classical Bayesian derivation :

$$p(m_i|y_c) = \sum_{k=1}^{card(E)} p(m_i|y_c, e_k) \cdot p(e_k|y_c) \quad (4)$$

Finally, masks can be designed either in a soft fashion by taking $p(m_i|y_c)$ as the mask value, or in a hard fashion by thresholding this probability :

$$m_i = (p(m_i|y_c) \geq \delta) \quad (5)$$

where δ is a threshold in the interval $[0, 1]$. The complete architecture of our mask estimator is presented in figure 2.

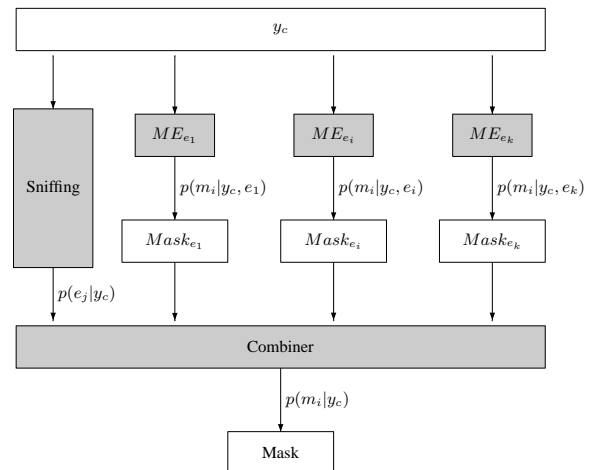


Fig. 2. Our missing data mask estimator architecture. “ ME_{e_k} ” is a mask estimator trained on the *a-priori* known environment e_k . “ $p(m_i|y, e_k)$ ” is the probability that the i^{th} coefficient of y is missing when e_k is supposed to be the corrupting environment. Each mask estimator ME_{e_k} provides an environment dependent mask “ $Mask_{e_k}$ ” from $p(m_i|y_c, e_k)$. These masks are combined according to the environment sniffing module that gives the probability of being in each *a-priori* noisy condition.

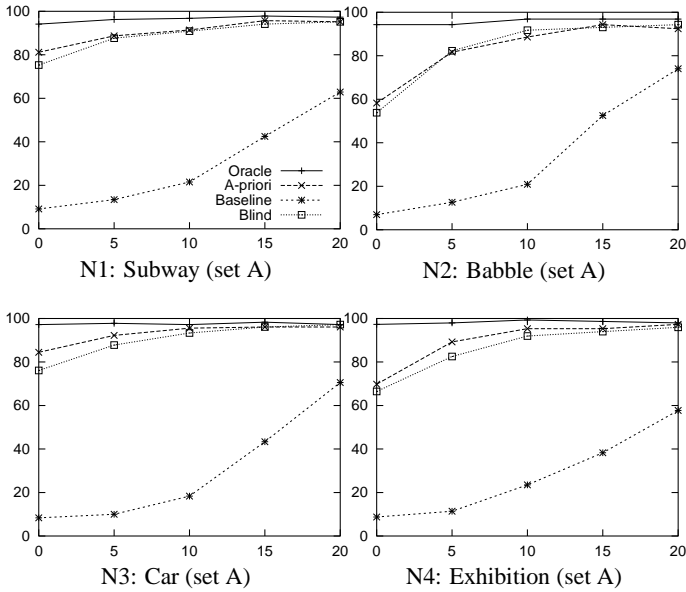


Fig. 3. Recognition accuracy for test set A of Aurora2 database.

4. EXPERIMENTAL SETUP

All the experiments reported in this paper are based on the Aurora2 database speaker independent connected digit recognition task [13].

4.1. Models training

The environment sniffing and mask models are trained on 800 sentences of the Aurora2 test sets and tested on 50 other sentences. Two parametrizations of the signal are used. The first one is the classical set of 13 Mel Cepstral Coefficients (MFCC) supplemented with their temporal derivatives, which gives a 26 dimensional feature vector. This parametrization is used to estimate the masks and the environments. The second one is the classical 32 Mel spectral coefficients with a cuberoot-compression. It is used for speech recognition and digits models training. The 28 environments (4 noise at 7 SNRs) of the training corpus of Aurora2 define our set of *a-priori* known environments. Two GMMs with 256 Gaussians are trained on each of these environments e_k , for each frequency band i . They respectively model the probability density functions $p(y_c|m_i, e_k)$ and $p(y_c|\bar{m}_i, e_k)$. In addition, a GMM with 256 Gaussians is trained on all the observations of a given environment. It models the probability density functions $p(e_k|y_c)$.

The *a-priori* probability that a spectrographic feature is missing $p(m_i|e_k)$ is a scalar value for each frequency band of each *a-priori* environment. Whole word digit models are trained on the Aurora2 clean speech training set. These models are standard 16 states HMMs with 7-component mixtures. The task grammar forces every sentence to begin and end by a silence.

4.2. Missing data recognition

Soft bounded marginalization is used during recognition. Any spectral coefficient is considered as missing when the local SNR is below 0 dB, and reliable otherwise. Thanks to cuberoot-compression, this

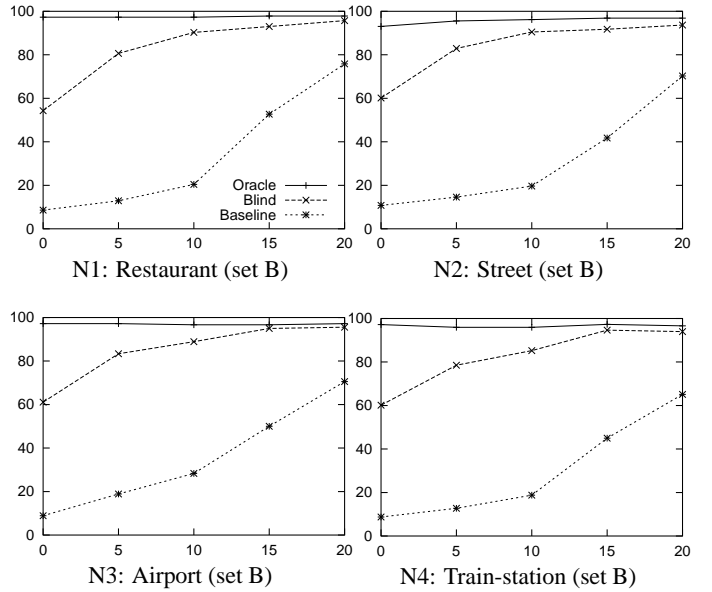


Fig. 4. Recognition accuracy for test set B of Aurora2 database.

implies that the speech contribution x_i of any observed missing coefficient y_i is in the interval:

$$0 \leq x_i \leq \sqrt[3]{y_i^3/2} \quad (6)$$

Conversely, the speech contribution x_i of any observed reliable coefficient y_i is in the interval:

$$\sqrt[3]{y_i^3/2} \leq x_i \leq y_i \quad (7)$$

For simplicity, we set $y_{i,snr0} = \sqrt[3]{y_i^3/2}$. Hence, the emission probability of a particular feature vector y is expressed as follow:

$$p(y|\Theta) = \prod_i \left\{ \frac{p(m_i|y_c)}{y_{i,snr0}} \cdot \int_0^{y_{i,snr0}} p(x_i|\Theta) dx_i + \frac{p(\bar{m}_i|y_c)}{y_i - y_{i,snr0}} \cdot \int_{y_{i,snr0}}^{y_i} p(x_i|\Theta) dx_i \right\} \quad (8)$$

where Θ reflects state parameters. This soft bounded marginalization was introduced in [14].

5. EXPERIMENTAL RESULTS

5.1. Validation of the approach with known environments

Figure 3 shows the recognition accuracy of the proposed approach on test set A. The “Baseline” and “Oracle” systems respectively give the recognition accuracy without masking any data, and with the optimal masks computed from the true SNR.

The “A-priori” system presents the results when the environment is known *a-priori*. We can observe that the proposed approach gives quite good recognition results, comparable with the other missing data systems reported in the literature.

The “Blind” system does not assume any knowledge about the environment. The results given by the “A-priori” and “Blind” systems are quite close, which validates the environment sniffing module. The average accuracy of the blind system on test set A is 81.47 %.

5.2. Tests with unknown environments

Figure 4 shows the recognition rates obtained on test set B, where the noise types are different from those used to train the mask and environment models. Only the blind system can be represented on these figures.

The average recognition accuracy of the blind system on test set B is 77.38 %. We can observe a slight decrease of the performances when there is a mismatch between the training and test environments. However, the recognition results are still comparable with the other state-of-the-art missing data recognition approaches. This confirms the potential of the proposed method.

6. DISCUSSION

The proposed approach is based on a set of known “typical” noisy environments. This raises two important issues: how can we deal with unknown environments that are completely different from the training conditions, and how large can be, or should be, the set of training environments.

Regarding the first issue, we are currently investigating some solutions to adapt the models to a new environment that do not belong to the training database.

About the second issue, a very large noisy database is probably not the best option, as the confusion between the GMMs might increase with the number of different environments. In this work, we considered 4 noises and 7 SNRs, which lead to 28 training conditions. The results show that such a number of environments is not excessive. If we want to use this system in many possible conditions, a solution may be to cluster all these environments hierarchically, and to use external algorithms to pre-select a small subset of them, before applying our combination method. An alternative to reduce the number of different environments would be to use a “canonical” noise database, which would represent concisely a wide variety of noises, as it is done in the *eigenvoices* method for speaker voices. Adaptation can further be considered to fine-tune the models to specific test conditions.

7. CONCLUSIONS AND FUTURE WORK

In this paper we have presented a new missing data mask estimator. Each spectrographic feature is classified as reliable or missing using an environment dependent model. The resulting masks, one for each candidate environment, are then combined. The noise dependent masks combination is led by an environment sniffing module that gives the probability to be in one of the training environments and at a given SNR. The current mask estimator defines a baseline and shall be modified in several aspects to improve its accuracy. In particular, the classical MFCC signal parametrization is used in our system, but better parameters can be used to distinguish reliable features from missing ones. Those proposed by Seltzer are good examples. Moreover, we plan to integrate temporal dependencies in our models, either in the form of HMMs, or with context dependent models similar to the n-gram models used in language modeling. Finally, different combination schemes can be studied.

8. ACKNOWLEDGEMENTS

This work was supported by the HIWIRE (Human Input That Works In Real Environments) project consortium.

9. REFERENCES

- [1] Andrew Morris, Jon Barker, and Herv Bourlard, “From missing data to maybe useful data: soft data modelling for noise robust ASR,” in *Proc. WISP-01*, Stratford-upon-Avon, England, April 2001, pp. 153–164.
- [2] Jon Barker, Phil Green, and Martin Cooke, “Linking auditory scene analysis and robust ASR by missing data techniques,” in *Proc. WISP*, Stratford-upon-Avon, England, April 2-3 2001, pp. 295–307.
- [3] Andrzej Drygajlo and Mounir El-Maliki, “Speaker verification in noisy environments with combined spectral subtraction and missing feature theory,” in *ICASSP*, Seattle, 1998, pp. 121–124.
- [4] Philippe Renevey, *Speech recognition in noisy conditions using missing feature approach*, Ph.D. thesis, Ecole Polytechnique Fédérale de Lausanne, 2001.
- [5] Jon Barker, Martin Cooke, and Phil Green, “Robust asr based on clean speech models: an evaluation of missing data techniques for connected digit recognition in noise,” in *Proc. EUROSPEECH*, Denmark, September 2001.
- [6] Hugo Van hamme, “Robust speech recognition using cepstral domain missing data techniques and noisy masks,” in *ICASSP*, Montreal, Quebec, Canada, 2004, vol. 1, pp. 213–216.
- [7] G. J. Brown, J. Barker, and D. L. Wang, “A neural oscillator sound separator for missing data speech recognition,” in *Proc. IJCNN-01*, Washington, DC, USA, 2001, vol. 4, pp. 2907–2912.
- [8] Sam T. Roweis, “Factorial models and refiltering for speech separation and denoising,” in *Proc. EUROSPEECH*, Geneva, Switzerland, 2003, pp. 1009–1012.
- [9] Jon Baker, Martin Cooke, and Daniel P.W. Ellis, “Decoding speech in the presence of other sources,” *Speech Communication*, vol. 45, no. 1, pp. 5–25, January 2005.
- [10] Michael L. Seltzer, “Automatic detection of corrupt spectrographic features for robust speech recognition,” M.S. thesis, Departement of Electrical and Computer Engineering, Carnegie Mellon University, 2000.
- [11] Wooil Kim, Richard M. Stern, and Hanseok Ko, “Environment-independent mask estimation for missing-feature reconstruction,” in *Proc. INTERSPEECH*, Lisbon, Portugal, September 2005, pp. 2637–2640.
- [12] Murat Akbacak and John H.L. Hansen, “Environmental sniffing : robust digit recognition for an in-vehicle environment,” in *Proc. EUROSPEECH*, Geneva, Switzerland, September 2003, pp. 1–4.
- [13] D. Pearce and H.-G. Hirsch, “The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions,” in *Proc. ICSLP*, Beijing, China, October 2000, vol. 4, pp. 29–32.
- [14] Andrew Morris, “Data utility modelling for mismatch reduction,” in *Proc. CRAC (workshop on Consistent & Reliable Acoustic Cues for sound analysis)*, Aalborg, Denmark, September 2001.