Missing data mask models with global frequency and temporal constraints

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

Missing data recognition has been developed in order to increase noise robustness in automatic speech recognition. Many different factors, including the speech decoding process itself, shall be considered to locate the masks. In this work, we are considering Bayesian models of the masks, where every spectral feature is classified as reliable or masked, and is independent from the rest of the signal.

This classification strategy can produce unrelated small "spots", while experiments suggest that oracle reliable and unreliable features tend to be clustered into time-frequency blocks. We call this undesired effect: the "checkerboard" effect.

In this paper, we propose a new Bayesian missing data classifier that integrates frequency and temporal constraints in order to reduce, or avoid, this "checkerboard" effect. The proposed classifier is evaluated on the Aurora2 connected digit corpora. Integrating such constraints in the missing data classification leads to significant improvements in recognition accuracy.

Index Terms: missing data recognition, mask estimation, temporal constraints, frequency constraints.

1. Introduction

The presence of background noise typically causes mismatches between training and testing conditions, which can significantly degrade the recognition accuracy of speech recognition systems. Many methods dealing with this problem have been developed over the last decades, such as noise reduction or speech model adaptation among others. Nevertheless, these techniques are not completely satisfying because most of them are based on estimates of the corrupting noise. Hence, they are still vulnerable to the effects of time varying noise such as music or concurrent speech.

Nowadays, 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 [1]. A number of experiments have demonstrated the potential of MFT [2]. The process of estimating such regions is called missing data mask estimation. It depends on many different factors, such as the location of speech formants, the characteristics of noise, or the discriminant power of speech models. It can be realized independently from the speech decoding process, or it can be fully integrated with it.

Several missing data mask estimation methods have been proposed over the last few years. Among them, Bayesian mask estimation has emerged. This technique consists in training statistical models for reliable and unreliable features and in estimating the most probable mask in the Baye's sense according to the features. This paper is focused on this approach.

We propose to take into account frequency and temporal constraints in the Bayesian mask estimation process. A classical approach to achieve this is to include into the observation vector its context, i.e. the spectro-temporal features that are close to it. However, this solution is only local. We propose here to take into account global constraints, which means that spectro-temporal features that are far from the coefficient of interest might influence the decision of the mask estimation process.

The organization of the paper is as follows. In section 2, we present the missing data recognition strategy retained. We then motivate and expose the frequency and temporal constraints integration in the missing data mask estimation in section 3. Section 4 presents experiments and validates the new approach.

2. Missing data recognition

.

Two different approaches to handle missing data during recognition can be used. The first one, called *marginalisation*, marginalises the observation likelihood of the corrupted (unreliable) features. The second one, called *data imputation*, estimates the contribution of speech to the unreliable features. Comparative tests with *marginalisation* and *data imputation* have been done [3]. The missing data strategy used in our work is the missing data marginalisation described below.

2.1. Static features marginalisation

In this work, we consider features that belong to the spectral domain. The spectral power is usually compressed by a function $\gamma(\cdot)$. Any static spectral coefficient is considered as missing when its local SNR is below 0 dB, and reliable otherwise. Hence, according to noise additivity in the spectral domain, the speech contribution x_i of any masked observation coefficient y_i lies in the interval:

$$0 \leq x_i \leq \gamma(\gamma^{-1}(y_i)/2) \tag{1}$$

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

$$\gamma(\gamma^{-1}(y_i)/2) \leq x_i \leq y_i \tag{2}$$

For simplicity, we set $y_{i,snr0} = \gamma(\gamma^{-1}(y_i)/2)$. Hence, the emission probability of a particular static feature vector y is expressed as follows:

$$p(y|\Theta) = \prod_i \left\{ \frac{m_i}{y_{i,snr0}} . \int_0^{y_{i,snr0}} p(x_i|\Theta) dx_i + \frac{1-m_i}{y_i - y_{i,snr0}} . \int_{y_{i,snr0}}^{y_i} p(x_i|\Theta) dx_i \right\} \tag{3}$$

where Θ reflects speech models state parameters, and $m_i = 1$ iff y_i is missing, $m_i = 0$ otherwise. This soft bounded marginalization was introduced in [4].

2.2. Dynamic features marginalisation

Dynamic features are computed from their static counter-parts. The following equation is often used for this purpose:

$$\Delta y_i(t) = \frac{\sum_{j=-N}^{j=N} j.y_i(t+j)}{\sum_{j=-N}^{j=N} j^2} \quad (4)$$

It is proposed in [5] to consider that a dynamic feature $\Delta x_i(t)$ is missing iff any static feature $x_i(t+j)$ is missing, with $j = -N \dots N$.

Since the missing data mask does not provide any information on the true value of the dynamic features, a full marginalisation is realized and the emission probability of a particular dynamic feature vector y is expressed as follows:

$$p(\Delta y(t)|\Theta) = \prod_i \left\{ m_i + (1 - m_i).p(\Delta y_i(t)|\Theta) \right\} \quad (5)$$

3. A new Bayesian missing data classifier

3.1. Motivations: the "checkerboard" effect

Seltzer [6] proposed a Bayesian classifier to label spectrographic features , 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.* [7] suggested a new training method based on subbands of colored noise. In one of our latest works [8], we presented a mask estimator that infers missing data masks by combining noise dependent masks. Each of them is generated by a Bayesian classifier trained on a typical noisy environment and the combination is led by an environmental sniffing module that estimates the probability of being in each training environment.

These methods have been developed in order to improve the robustness of the classifiers to unknown environments. They are based on diagonal-covariance Gaussian Mixture Models (GMMs), and thus estimate missing data masks for each spectrographic feature independently of the others. We believe that estimating feature reliability independently from its context can lead to unrealistic mask. Indeed, this procedure might build masks that are composed of many unrelated small "spots". This problem has been evoked in [6]. We propose here to integrate temporal and frequency constraints in order to limit this "checkerboard" effect.

3.2. Frequency constraints

Given an acoustic vector $y(t)$ of dimension N, there are 2^N possible masks $m(t) = (m_1(t), m_2(t), \dots, m_N(t))$ associated to $y(t)$, with $m_i(t) \in \{0, 1\}$. Because of this combinatory explosion, it is impossible to directly train every possible mask for a realistic number of frequency bands. However, the actual distribution of oracle masks in this large space is sparse and probably concentrated around localized regions. This intuitively may come from the fact that masks depend both on the noise and speech characteristics: for instance, high energetic frequencies of speech are rarely masked while low energetic frequencies are often masked. The masks thus tend to follow the same structures and to concentrate on well-defined parts of the search space.

In order to integrate frequency constraints, we propose to train a GMM for each of the K most frequent masks. K is chosen so that the cumulative occurences of these masks reach a covering threshold α. Each mask that does not belong to these K most frequent is replaced by the closest one according to the euclidian distance in the mask space.

α	0.55	0.60	0.65	0.70	0.75
number of masks	2	4	7	12	18
oracle	96.69	96.69	96.69	96.69	96.69
projected oracle	81.65	84.72	92.19	94.42	94.65
α	0.80	0.85	0.90	0.95	1.00
number of masks	31	54	109	244	4096
oracle	96.69	96.69	96.69	96.69	96.69
projected oracle	95.93	96.07	96.36	96.55	96.69

Table 1: *Recognition accuracy and number of masks in function of the covering threshold α.*

An experiment was conducted to evaluate the decrease in performances induced by this approach. For a set of thresholds α, we have computed the number of remaining masks, the recognition accuracy with true oracle masks and with oracle masks projected on the K most frequent ones. We use acoustic vectors with 12 frequency bands that lead to 4096 potential masks. This test was realized on a part of the Aurora2 multicondition training corpora. Results are summarized in table 1.

Note that only 31 masks among the 4096 potential ones cover 80% of the oracle masks. Moreover, projecting the 4065 other masks onto this small subset of 31 masks only slightly degrades the recognition accuracy from 96.69% down to 95.93%. This threshold, which has been estimated on the training corpus, will be used in the following experiments.

3.3. Temporal constraints

We have chosen to integrate temporal constraints into the mask estimation procedure by applying transition probabilities between subsequent masks. These transitions are defined locally, but their influence is global as the chosen masks belong to the best path that maximizes the observation likelihood over the whole sentence. Concretely, we have built an ergodic Hidden Markov Model (HMM), where each state contains one GMM mask model. The HMM thus contains K states and K^2 transitions, which are trained separately on the oracle masks computed on the training corpus, and eventually projected onto the reduced K-masks set.

4. Experiments

.

All the experiments reported in this paper are based on the Aurora2 database speaker independent connected digit recognition task [9].

The missing data models are trained on the Aurora2 multicondition training corpora. The models feature domain is the classical MFCC. 12 coefficients are used with the energy value and supplemented with their temporal derivatives, which gives a 26 dimensional feature vector. Finally a CMN normalisation is applied. 512 Gaussians with diagonal covariance matrix are trained for each GMM and each HMM state.

The feature domain of the speech acoustic models, which is also the marginalisation domain of missing data, is the 12-bands Mel spectral domain, with cube-root compression of the speech power. 12 temporal derivatives are further added, leading to a 24 dimensional feature vector. The masking scheme is hard, which

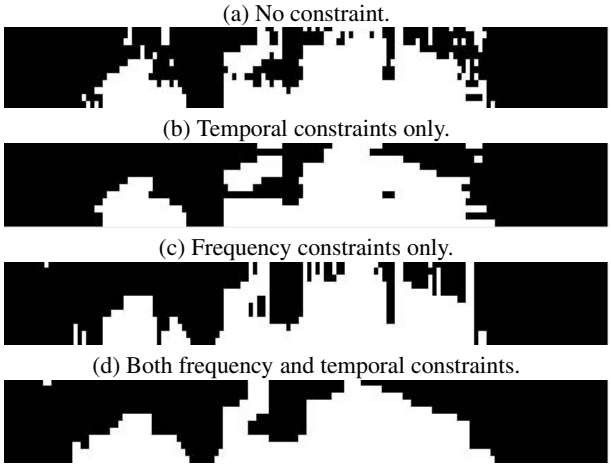

Figure 1: *Comparison between masks generated by (a) the unconstrained mask models, (b) the HMM, (c) the GMM with quantized masks and (d) the HMM with quantized masks.*

means that the mask associated to every coefficient of this feature vector is binary. Training and testing the acoustic models is realized by the HTK scripts of the Aurora2 database. HTK has been modified to perform the marginalization procedure described in section 2.

Four experiments are presented next, to evaluate the impact of the temporal and frequency constraints on the recognition accuracy. Note that in every experiment, mask GMM models are trained on the whole MFCC frame, even though they model the mask for a single spectral coefficient.

4.1. Unconstrained system

For each frequency band i two GMMs are trained: $M_{i,u}$ for unreliable and $M_{i,r}$ for reliable data. This missing data model architecture does not take into account neither frequency nor temporal constraints. It is called hereafter "unconstrained system". The mask of each coefficient is infered as follows:

$$m_i(t) = \arg\max_{M_i}(p(M_i|y_i(t))) \tag{6}$$

with $M_i = (M_{i,u}, M_{i,r})$.

4.2. Integrating temporal constraints only

In this system, a single mask model is built for each frequency band. It is composed of a 2-states ergodic HMM: the first one contains the $M_{i,u}$ GMM, and the second one the $M_{i,r}$ GMM. Then, the transition probabilities between reliable and unreliable data are trained independently.

Let $S_i = (m_i(1), \cdots, m_i(T))$ denote a HMM state sequence for a given frequency band i. The missing data mask for this frequency band is given by the best state sequence S_i^{best}:

$$S_i^{best} = \arg\max_{S_i}(p(S_i|y_i(t), \Theta)) \tag{7}$$

where Θ reflects the HMM parameters.

4.3. Integrating frequency constraints only

In this system, K full-band GMMs $M_1, M_2, \ldots, M_K$ are trained on the training frames that are aligned with the most frequent oracle masks. Now, each GMM models a complete mask frame, and not a single mask coefficient, as it was previously the case. Hence, we do not any more distinguish reliable from unreliable GMMs, as any frame mask may include both masked and unmasked coefficients.

The α oracle covering threshold is set to 0.80, which leads to 31 most frequent frame masks and 31 GMMs. This threshold seems to be a good compromise between models confusion and loss of recognition accuracy as mentioned in the previous section. During testing, the mask associated to each frame is computed as follows:

$$m(t) = \arg\max_{M}(p(M|y(t))) \tag{8}$$

where $M \in \{M_1, M_2, \ldots, M_K\}$.

4.4. Integrating both frequency and temporal constraints

The last mask model includes both frequency and temporal constraints. The missing data classifier is composed of a 31 states ergodic HMM, where each state respectively contains one of the K GMMs defined in section 4.3.

Let $S = (m(1), \cdots, m(T))$ denote a HMM state sequence. The missing data mask for this frequency band is given by the best state sequence S^{best}:

$$S^{best} = \arg\max_{S}(p(S|y_f(t), \Theta)) \tag{9}$$

where Θ reflects the HMM parameters.

4.5. Results

Figure 1 illustrates the impact of integrating temporal and frequency constraints during missing data classification. Missing data are presented in black while reliable data are presented in white. As mentioned in section 3.1, estimating missing data independently leads to the checkerboard effect shown in figure 1(a). Applying temporal constraints only smooths masks along the time axis and can produce discontinuities along frequency axis, (fig. 1(b)) while applying frequency constraints smooths masks along frequency axis and can produce discontinuities along time axis(fig. 1(c)). Applying both constraints tends to cluster data in homogeneous blocks without discontinuities (fig. 1(d)) and thus reduces the resulting checkerboard effect.

Figure 2 presents digit recognition accuracy for the four proposed systems and the HTK baseline systems. The digit accuracies are averaged over all noisy conditions of each test set at a given SNR. The proposed constraints improve the recognition accuracy, and the best performances are obtained when combining both of them.

The improvement due to frequency constraints is greater for test set B than for test set A. The same types of noise are used both in multicondition training and in test set A, while test set B defines completely new environments. Therefore, frequency constraints seem to increase the robustness of missing data to unknown environments, at least for moderate environmental mismatch, while temporal constraints improve performances in all cases. It can be noted that frequency constraints slightly degrade results for test set

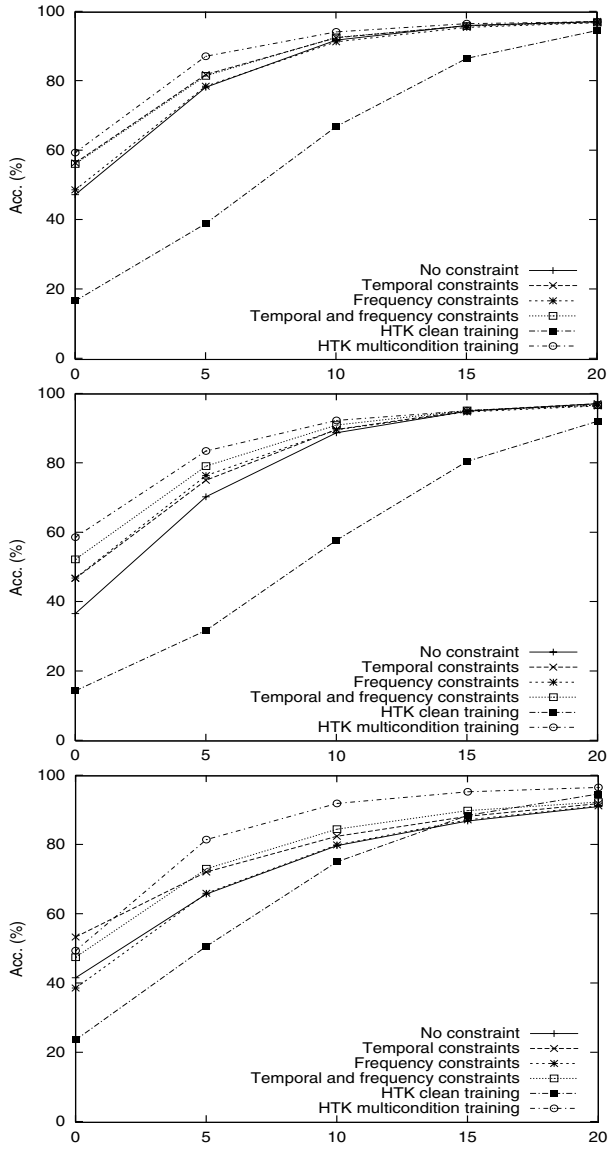

Figure 2: *Digit recognition accuracy of the four proposed missing data systems, and the baseline HTK systems, on the Aurora2 test sets A (top) ,B (middle) and C (bottom) at a SNR from 0 to 20 dB.*

C because it exhibits a different frequency characteritic compared to test sets A and B.

Averaged recognition performances from 20 dB to 0 dB on the three Aurora2 test sets are shown in table 2.

It can be noted that only 12 frequency bands are used for these tests and that masks are binary. In spite of this poor mask design, performances are good. Investigations to increase the number of frequency bands have been carried out. However, increasing the number of frequency bands leads to an explosion of the number K of retained masks. For example, 2826 remaining masks are obtained with an oracle covering threshold $\alpha = 80\%$ using 24 frequency bands among the 16 777 216 possible ones. The number of retained masks can be reduced to 70 with $\alpha = 60\%$, but this low threshold might lead to poor mask models. This important increase of representative masks for higher parametrization dimensionality appears to be the main limitation of the proposed method.

	set A	set B	set C
HTK clean training	61.13	55.57	66.68
No constraint	82.27	77.66	73.14
Frequency constraints	82.47	81.01	72.76
Temporal constraints	84.44	82.23	80.17
Both constraints	84.94	83.04	77.62
HTK multicondition training	87.29	85.52	83.12

Table 2: *Averaged recognition accuracy for the 3 Aurora2 test sets.*

5. Conclusions

Bayesian modelling of individual time-frequency masks might result in a dispersion of the masks on the spectrogram, which we call the checkerboard effect. We have proposed in this work to reduce this effect by incorporating frequency and temporal constraints in the missing data classification process. The proposed solutions have been evaluated on the standard Aurora2 database, and experimental results show the effectiveness of the proposed approach: the checkerboard effect is significantly reduced, and recognition accuracy is improved. This approach shall be adapted next to support more frequency bands in the parameterization module.

6. Acknowledgements

This work was partly supported by the IST HIWIRE (Human Input That Works In Real Environments) project (contract number 507943) of the sixth framework program supported by the european commission.

7. 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, 2001.

[2] J. Barker, P. Green, and M. Cooke, "Linking auditory scene analysis and robust ASR by missing data techniques," in *Proc. WISP*, Stratford-upon-Avon, England, 2001.

[3] A. Morris, J. Barker, and H. Bourlard, "From missing data to maybe useful data: soft data modelling for noise robust ASR," in *Proc. WISP*, Stratford-upon-Avon, England, 2001.

[4] A. Morris, "Data utility modelling for mismatch reduction," in *Proc. CRAC (workshop on Consistent & Reliable Acoustic Cues for sound analysis)*, Aalborg, Denmark, 2001.

[5] J. Barker, L. Josifovski, M. Cooke, and P. Green, "Soft decisions in missing data techniques for robust automatic speech recognition," in *Proc. ICSLP*, Beijing, China, 2000.

[6] M. 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.

[7] W. Kim, R. M. Stern, and H. Ko, "Environment-independent mask estimation for missing-feature reconstruction," in *Proc. INTERSPEECH*, Lisbon, Portugal, 2005.

[8] S. Demange, C. Cerisara, and J.-P. Haton, "Mask estimation for missing data recognition using background noise sniffing," in *Proc. ICASSP*, Toulouse, France, 2006.

[9] 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, 2000.