Towards Missing Data Recognition with Cepstral Features

Christophe Cerisara

LORIA - UMR 7503
54500 Vandoeuvre-les-Nancy - FRANCE
cerisara@loria.fr

Abstract

We study in this work the Missing Data Recognition (MDR) framework applied to a large vocabulary continuous speech recognition (LVCSR) task with cepstral models when the speech signal is corrupted by musical noise. We do not propose a full system that solves this difficult problem, but we rather present some of the issues involved and study some possible solutions to them. We focus in this work on the issues concerning the application of masks to cepstral models. We further identify possible errors and study how some of them affect the performances of the system.

1. Introduction

The two main issues to solve when using missing data recognition are:

- To find the masks that represent the corrupted or unreliable spectro-temporal regions;
- To compute the likelihoods that the speech models have generated the acoustic observations, given the fact that some spectro-temporal regions are masked.

In this work, we will not consider the former problem at all. We rather focus on the latter issue. However, the fact that the masks must segregate reliable and corrupted coefficients implies that a frequency-localized noise must also be localized in the feature domain. This is certainly not the case for cepstral coefficients, and this is why most of the previous work in the field of MDR has been realized with "spectrographic-like" models [1, 2]. On the other hand, it is well-known that MFCCs are less sensitive to amplitude mismatch than spectrographic coefficients, and are also more robust to noise. We study here the possible use of MDR techniques with MFCC models. Our other main contributions concern the use of MDR in a large vocabulary continuous speech recognition task where the speech signal is contaminated by musical noise.

2. Missing data recognition with cepstral models

2.1. General considerations

There are two main solutions to solve the second issue mentioned in section 1:

- *Data imputation*: The system tries to estimate somehow the unknown value of the "corrupted" observations;
- *Marginalization*: The recognizer marginalizes the likelihoods over the unknown observations, which is an optimal decision given the masks in the Bayes' sense [3].

The best results are reported in the literature for the marginalization process [1], but this technique can only be applied to spectral features at a reasonable complexity. On the other hand, data imputation can more easily be applied to cepstral models, as once the noisy observations have been replaced by some estimated uncorrupted values, further processing of these vectors (such as a discrete cosine transform) can be realized and recognition in the cepstral domain can thus be performed. This is why we have chosen to describe our approach within the data imputation framework.

A few other works such as [4, 5] deals with the use of MFCC models within the MDR framework. We extend here such studies by (i) proposing and testing a new system that uses MFCC models in a LVCSR task corrupted by background music and (ii) identifying and studying specific recognition errors that may result from the use of cepstral features.

2.2. Class-conditional imputation

In the following, the upper index ^{l} represents log-spectral vectors. When ^{l} is omitted, MFCC vectors are considered. Let X^l be the observation at time t. We decompose

$$X^l = \begin{bmatrix} X^l_{|M} \\ X^l_{|\bar{M}} \end{bmatrix} \quad (1)$$

where $X^l_{|M}$ and $X^l_{|\bar{M}}$ respectively represent the masked and unmasked coefficients of X^l.

Class-conditional imputation [2] (or *mean imputation*) simply consists to impute $\hat{X}^l_{|M} = \mu^l_{|M}$ where $\mu^l_{|M}$ represents the masked coefficients of the mean of the Gaussian model aligned with X^l. This technique assumes that $\hat{X}^l_{|M}$ is independent of $X^l_{|\bar{M}}$. It is reported in [1] that using the correlation between the masked and unmasked coefficients improves the results, but at the cost of a much higher complexity. We thus assume in the next sections that $X^l_{|M}$ and $X^l_{|\bar{M}}$ are independent and we propose some improvements of this method.

2.3. Vocabulary size

Very good results are reported for small-size vocabulary MDR tasks [6]. A few other works deal with medium-size vocabulary [3, 2], but the results seem to be less good. To our knowledge, there is no result reported for large vocabulary, and we propose here to study some MDR techniques in a LVCSR task.

2.4. Baseline algorithm

We assume in the next experiments that the masks are given. Such masks are usually called *oracle* masks [2]. This approach does not give a clear understanding of how MDR might perform in real recognition tasks, but it is very useful to isolate the

problems when we study a new approach in the field of MDR. We should then keep in mind that in a more realistic use of our system, these masks will have to be estimated and the performances will probably decrease.

The algorithm we use is based on the classical Viterbi algorithm, modified as follows:

- For a given alignment between the (static) cepstral frame X and the (static) cepstral clean speech distribution $S \sim N(\mu, \Sigma)$, both vectors X and $E[S]$ are transformed into the log-power-spectral domain: For X, this simply consists to compute $X^l = C^{-1} \cdot X$ where C is the discrete cosine matrix. For $E[S]$, we know that:

$$C^{-1} \cdot S \sim N(C^{-1}\mu, C^{-1} \cdot \Sigma \cdot C^{-T}) \quad (2)$$

  Therefore, $E[S^l] = \mu^l = C^{-1}\mu$.
- $X^l_{|M}$ is replaced by:

$$\hat{X}^l_{|M} = \left\{ \begin{array}{l} \mu^l_{|M} \text{ if } X^l_{|M} > \mu^l_{|M} \\ X^l_{|M} \text{ if } X^l_{|M} < \mu^l_{|M} \end{array} \right\} \quad (3)$$

- The unmasked coefficients $X^l_{|\bar{M}}$ are not modified:

$$\hat{X}^l_{|\bar{M}} = X^l_{|\bar{M}} \quad (4)$$

  The imputed observation vector is now

$$\hat{X}^l = \left[\begin{array}{c} \hat{X}^l_{|M} \\ \hat{X}^l_{|\bar{M}} \end{array} \right] \quad (5)$$

- This vector is then transformed back into the cepstral domain:

$$\hat{X} = C \cdot \hat{X}^l \quad (6)$$

- Dynamic cepstral parameters are usually computed as a linear combination of static parameters of successive frames such as [1]:

$$\Delta X(t) = 2X(t+2) + X(t+1) - X(t-1) - 2X(t-2) \quad (7)$$

  In the speech models, only $\Delta\mu$ is known (and not the individual terms that appear in the right-hand side of equation 7). To impute $\Delta X(t)$, we thus have to assume that the mask defined at time t is also valid at times $t-2$, $t-1$, $t+1$ and $t+2$. Therefore, the same coefficients of $X(t-2)$, $X(t-1)$, $X(t+1)$ and $X(t+2)$ are masked and we can decompose

$$\Delta X^l(t) = \left[\begin{array}{c} \Delta X^l_{|M}(t) \\ \Delta X^l_{|\bar{M}}(t) \end{array} \right] \quad (8)$$

  based on the mask defined at time t. *Unbounded* imputation can then be applied to this vector:

$$\Delta \hat{X}^l(t) = \left[\begin{array}{c} \Delta\mu^l_{|M} \\ \Delta X^l_{|\bar{M}}(t) \end{array} \right] \quad (9)$$

  The same principle is used to impute $\Delta^2 X(t)$.
- The emission likelihood is computed with this modified observation and the Viterbi algorithm repeats this procedure for every possible alignment.

The complexity of this algorithm can be made minimal by storing in memory a log-spectral version of the speech models and by using a front-end that directly computes the observations in the log-spectral domain before the DCT.

[1] Notation: t is omitted for static vectors but is kept in this derivation concerning the dynamic parameters

2.5. Improvement of the algorithm

The *bounded imputation* technique [2] is an extension of the imputation procedure presented in 2.2 and is used in the algorithm described in section 2.4. Its idea consists to replace the observation $X^l_{|M}$ by the model's mean log-power $\mu^l_{|M}$ if and only if $X^l_{|M} > \mu^l_{|M}$. This is reasonable because the energy of the speech plus noise should be greater than the energy of the speech alone. We propose here to further extend this principle by considering that when $X^l_{|M}$ is close to $\mu^l_{|M}$, then it is unlikely that the noise has a great influence on the recognition accuracy. Replacing these masked values by some approximations may be worse than doing nothing. Therefore, we decide to replace $X^l_{|M}$ by $\mu^l_{|M}$ iff $X^l_{|M} > \alpha\mu^l_{|M}$, where α is estimated on a development corpus.

On the other hand, preliminary experiments suggested that even when masking is really needed, it may actually be better to impute an intermediate value between $X^l_{|M}$ and $\mu^l_{|M}$. We thus propose to replace $X^l_{|M}$ by $\beta\mu^l_{|M}$ rather than by $\mu^l_{|M}$.

To summarize, equation 3 becomes:

$$\hat{X}^l_{|M} = \left\{ \begin{array}{l} \beta\mu^l_{|M} \text{ if } X^l_{|M} > \alpha\mu^l_{|M} \\ X^l_{|M} \text{ if } X^l_{|M} < \alpha\mu^l_{|M} \end{array} \right\} \quad (10)$$

The role of α is to reduce the number of coefficients effectively masked by adjusting the threshold used in the "bounded" paradigm, while the role of β is to "soften" the effect of masking by finding a compromise between completely masking the observation and not masking it at all.

3. First experimental results

3.1. Experimental setup

Experiments are realized on the BREF80 evaluation task [7]. It is a large vocabulary (20000 words) continuous speech recognition task in French, quite similar to the English Wall Street Journal task. 40 french monophones are modeled with 3 states left-to-right HMMs. We have adapted the *julius* recognition engine [8] to handle missing data as explained above. This engine makes use of two passes with a words bigram in the first pass and a trigram in the second pass. 39 MFCC parameters (included c0, delta and acceleration) are computed every 10 ms. 24 filters are used in the log-spectral domain. 300 sentences pronounced by 20 speakers are used for testing and 2000 sentences pronounced by 70 speakers for training. All the test sentences are "corrupted" by a background music, in that case Bach's *Chaccone*. The oracle masks are computed by comparing the corrupted and clean log-spectral observations: when the difference between them is greater than a given threshold, then the coefficient is masked.

3.2. Bounded factors

The first experiments reported here concern the estimation of the factor α described in 2.5. Figure 1 gives the word error rate in function of α on the 10 sentences of the development corpus.

We can observe that, as supposed, some improvement may be obtained by enforcing the "bounded" paradigm, i.e. by slightly increasing the masking / nomasking threshold. Note that, due to the size of the development corpus, the difference in Word Error Rate (WER) is not significant. But as the systems and corpus are exactly the same and only differ in α, we still choose α based on this experiment.

For the next experiments, we set $\alpha = 1.1$. We then evaluate the "optimal" value of β on the same development corpus. The

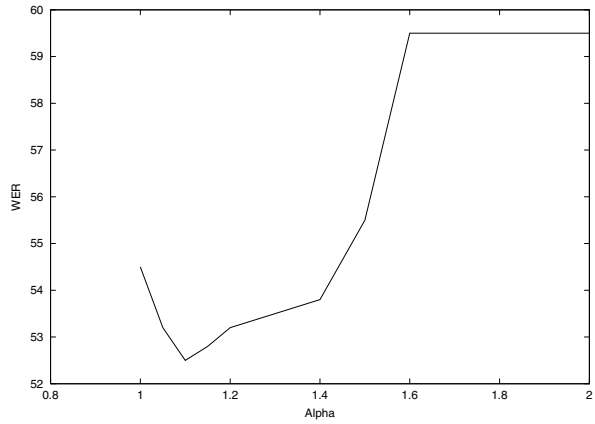

Figure 1: *Influence of α on the WER (development corpus).*

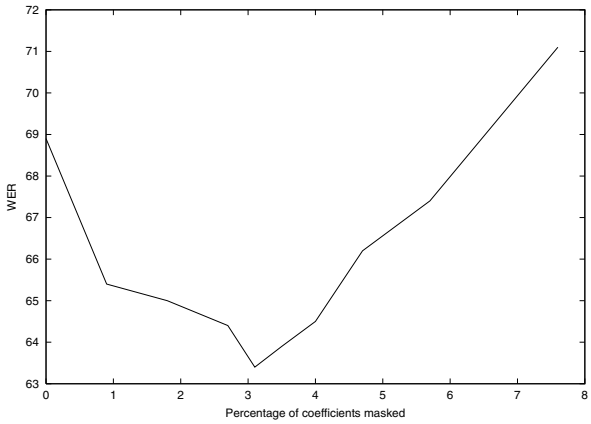

Figure 3: *Influence of the number of masks on the WER (test corpus).*

WER is represented in figure 2 in function of β. Note that when $\beta = 0$, the system behaves like a classical recognition engine without missing data, while when $\beta = 1$, the basic bounded missing data algorithm is used without any modification.

It is interesting to observe that the "optimal" value of β is 0.6, which is far from the basic imputation scheme obtained with $\beta = 1.0$. This supports the idea that it might be better to find a good "compromise" between the missing data paradigm and the classical one, than to simply use one or the other. This is also a possible interpretation of the "soft mask" procedure described in [1].

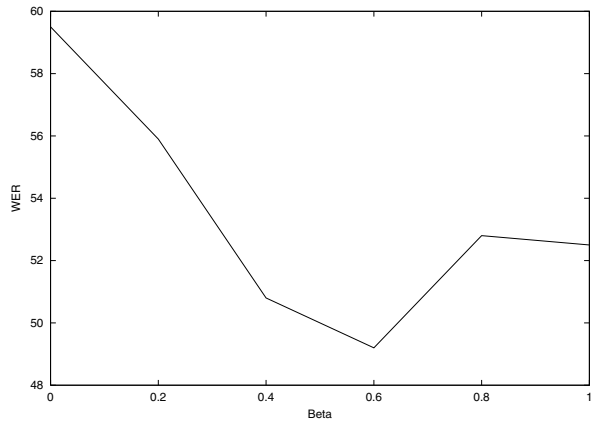

Figure 2: *Influence of β on the WER (development corpus).*

3.3. Number of masks

The next experiment is realized with $\alpha = 1.1$ and $\beta = 0.6$ on the whole test set (300 sentences). Its role is to evaluate the influence of the number of coefficients masked on the recognition accuracy.

Figure 3 gives the word error rate in function of the number of masked coefficients (this number is increased by decreasing the threshold used to compute the oracle masks). This number is computed after the test described in section 2.5 and represents the ratio of imputed coefficients in all the Viterbi paths.

As expected, when the number of masks increases too much, the WER increases dramatically. But it is surprising than the optimal number of imputed coefficients is only about 3 %, which is very few. As discussed in [1], part of the explanation may be related to the difficulty to have clearly discriminant acoustic models in a LVCSR task. We investigate in next sections other possible explanations.

3.4. Possible causes of recognition errors

For the baseline system (not MDR), the errors originate from:
(i) The mismatch between the corrupted observations and the clean models.
For MDR recognition, the errors originate from:
(i) The mismatch between the corrupted observations that are not masked and the clean models;
(ii) The lost of acoustic information in the masked coefficients (or equivalently the mistakes realized when imputing the missing coefficients).

The more masks are used, the more errors (ii) and the less errors (i) occur. The difficulty of MDR consists to find a good compromise between both kinds of errors. We study separately each kind of error in section 4.

4. A study of MDR errors

4.1. Study of the cepstral filtering effect

One of the issues in the case of cepstral parameterization is related to the filtering effect of the DCT. Indeed, as the number of cepstral parameters is usually not the same as the number of filterbanks in the log-spectral domain, some smoothing is introduced when transforming the imputed observations into the cepstral domain. This smoothing has two effects:

- It modifies the imputed values, that are not any more equal to the means of the models ;
- It modifies the unmasked values.

Both effects are not desired and may introduce new errors of types (i) and (ii). To test this issue, we have made some experiments by setting the number of filterbanks (N_{filt}) equal to the dimension of cepstral vectors $N_{cep} = 13$. Figure 4 compares $N_{filt} = 13$ with $N_{filt} = 24$.

We can note that the WER increases when $N_{filt} = 13$. this can be explained by the fact that the description of the observation in that case is less precise than with $N_{filt} = 24$. But more importantly, this shows that the smoothing effect introduced by

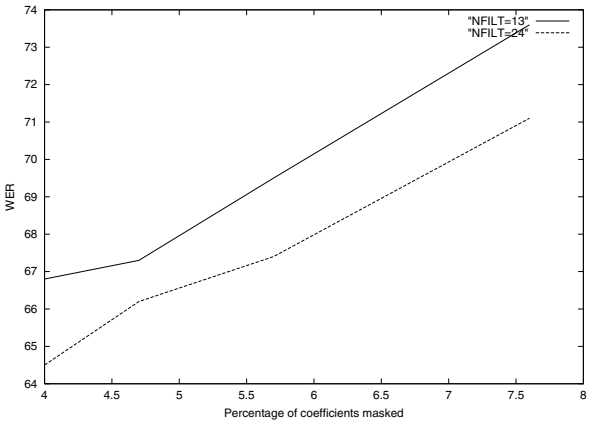

Figure 4: *Influence of the smoothing effect of the DCT on the WER.*

the DCT is not a problem and this validates the use of cepstral models in the framework of MDR.

4.2. Study of each kind of errors

In the next experiment, we try to remove all errors (i) in order to isolate errors (ii). To achieve this, just after the oracle masks are computed, the signal is processed by replacing every unmasked log-spectral coefficient by its uncorrupted (clean) value. Reversely, we also remove all errors (ii) by replacing every *masked* log-spectral coefficient by its clean value, and then realizing a classical (not MDR) recognition. This is equivalent at using an ideal imputation method.

Table 1 compares this system with the previous one and the baseline when 3.1% of the coefficients are masked.

Table 1: *Respective influence of each kind of errors on WER.*

system	WER
baseline	68.9 %
MDR: errors (i) & (ii)	63.4 %
MDR: only errors (ii)	52.4 %
MDR: only errors (i)	44.5 %
baseline without noise	36.7 %

We can note that both types of errors strongly affect the recognition rate. Consequently, we can imagine the following improvements for the system:

- The "oracle" masks are not that good, probably because the energy criterion on which they are based is not adequate. A better criterion would be to choose the masks that have a strong influence on the recognition accuracy. The corresponding spectro-temporal regions may have only a small energy mismatch due to the noise, and reversely other regions with an important mismatch may not really influence the WER. Such masks would certainly reduce errors (i).
- Better imputation procedures are needed to reduce errors (ii). Previous results show that reducing this kinds of errors may lead to an important improvement. Therefore, it might be worth to adapt solutions such as conditional imputation to cepstral models.

5. Conclusions

The original contributions of the work presented here are the following:

- Proposal of a system to test MDR in a LVCSR task corrupted by background music;
- Proposal and justification of a framework to use MDR with cepstral models;
- Extension of the "bounded imputation" procedure to (i) reduce the number of unneeded masks, and (ii) soften the effect of masking by adjusting the compromise between the MDR and classical frameworks;
- Analysis of the errors realized by a MDR system and experimental study of each type of error separately.

We believe the MDR paradigm might be more widely used if it becomes compliant with standard MFCC models. The main objective of this work is thus to create and strengthen a sane basis for MDR with cepstral models. But there are still many problems to solve before such solutions can be used in real-life situations. The most important is to adapt or create a better imputation procedure than the one presented here. Then, research objectives shall focus on the estimation of the masks.

6. References

[1] Cooke, M. and Green, P. and Josifovski, L. and Vizinho, A., "Robust automatic speech recognition with missing and unreliable acoustic data", Speech Communication, Vol. 34, N. 3, June 2001.

[2] Ramakrishnan, B. R., "Reconstruction of incomplete spectrograms for robust speech recognition", PhD. thesis, Department of Electrical and Computer Engineering, Carnegie Mellon University, April 2000.

[3] Morris, A. and Cooke, M. and Green, P., "Some solutions to the missing feature problem in data classification, with applications to noise robust ASR", in proceedings ICASSP, Seattle, Washington, USA, 1998, p. 737–740.

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

[5] Renevey, P., "Speech recognition in noisy conditions using missing feature approach", PhD. thesis, École Polytechnique Fédérale de Lausanne, 2000.

[6] Barker, J. and Cooke, M. and Ellis, D., "Decoding speech in the presence of other sound sources", in proceedings ICSLP, Beijing, China, 2000.

[7] Lamel, L. F., Gauvain, J. L. and Eskenazi, M., "BREF, a large vocabulary spoken corpus for French", in proceedings EUROSPEECH, 1991.

[8] Lee, A., Kawahara, T. and Shikano, K., "Julius – an open source real-time large vocabulary recognition engine", in proceedings EUROSPEECH, 2001, p. 1691–1694, http://winnie.kuis.kyoto-u.ac.jp/dictation/.