Exploiting confidence measures for missing data speech recognition

Christophe Cerisara

Exploiting confidence measures for missing data speech recognition

Christophe Cerisara
LORIA UMR 7503 - France

Automatic speech recognition in highly non-stationary noise, for instance with a competing speaker or background music, is an extremely challenging and still unsolved problem. Missing data recognition is a robust approach that is well adapted to this kind of noise. A standard missing data technique consists in marginalizing out, from the observation likelihoods computed during decoding, the contribution of the spectro-temporal fragments that are dominated by noise. However, such an approach can hardly be applied to advanced parameterization domains that do not separate speech from noise frequencies, such as the cepstrum or ETSI AFE. We propose in this work to extend this technique to such parameterization domains, and not only to spectrographic-like front-ends as it was the case before. This is realized by masking the observations that favor erroneous decoding paths, instead of masking the features that are dominated by noise. These new missing data "masks" are now estimated based on speech recognition confidence measures, which can be considered as indicators of the reliability of decoding paths. A first version of this robust algorithm is evaluated on the French broadcast news ESTER corpus.

1 Introduction

1.1 Robust automatic speech recognition

Robutsness of automatic speech recognition to noise is still a challenge for nowadays speech processing technologies. Many approaches have been proposed to address this issue. All these methods are classically classified into a few fundamental approaches:

- Denoising algorithms, which try to remove the contribution of noise in the signal before recognizing it;
- Robust parameterization algorithms, which try to exploit redondancy of speech information in the signal to encode only the speech contribution in the acoustic vectors;
- Adaptation algorithms, which modify the acoustic models so that they better match the test signal: intuitively, this can be interpreted as adding noise in the speech models.
- Uncertainty decoding algorithms, which modify the decoding process so that it takes into account the uncertainty in the observations that is due to noise. Missing data recognition belongs to this class of algorithms.

Missing data recognition is an interesting approach that is well suited to tackle nonstationary noise. Its main principles are described next.

1.2 Missing data recognition

Missing data recognition is based on the assumption that a given spectral band at a given time is dominated by a single acoustic source - either speech or noise. This information is encoded into *masks* that state whether each spectral coefficient is reliable or not for speech recognition.

Hence, there are two major problems in missing data recognition:

1. Estimating the masks;
2. Exploiting these masks in speech recognition.

Note that both problems can be jointly solved, such as in [2].

Many solutions exist to solve the first issue, which are reviewed in [5].

The second issue can be solved by two techniques, respectively data marginalization and data imputation. The former marginalizes out the contribution of masked coefficients during likelihood computation, while the latter replaces masked coefficients with an estimate of speech contribution only.

The main limitation of missing data recognition is that it requires parameterization domains in which the basic assumption used to discriminate speech and noise coefficients is valid, which is only true for a few front-ends, such as log-spectral, wavelet and auditory-based processing. This is not the case for most advanced front-ends.

2 Related work

Several previous works have proposed some solution to exploit missing data recognition techniques with better front-ends than the log-spectral domain, and in particular with cepstral features, such as [11], [14], [4], [16] and [15]. Most of these approaches basically propose to impute the missing observations in the cepstral domain when the masks are defined in the spectral domain.

Different front-ends can also be considered, such as auditory-based parameters [6], or the PROSPECT features [17], which are derived from the cepstrum but further include a second observation stream modeling the residual spectrum. They have proven to be quite efficient and well-adapted to missing data recognition.

The solution proposed next differs from the previous algorithms in several ways:

- The previous cepstral-based solutions rely on data imputation, because of the difficulty to derive data marginalization in cepstrum-like front-ends. The current work can handle both data imputation and marginalization without restrictions. It is applied next with marginalization.
- All previous solutions are dedicated to a specific parameterization domain, usually derived from the cepstral domain. The algorithm derived next is theoretically independent on the front-end, and may be applied with any kinds of features. This is particularly interesting, as robust parameterization domains tend to evolve very quickly: see for instance the ETSI AFE front-end [9], or the competing OGI Qualcomm features [1], or many other features, such as [8] for instance. The following method is the only one that supports all of these front-ends, at least theoretically.
- It is the first time a missing-data framework naturally supports dynamic coefficients, which were previously handled only by dedicated heuristics.

3 Front-End independent missing data recognition

3.1 Masks definition

In the context of automatic missing data recognition, masks can be defined as a function M that associates a value in a domain D to any speech coefficient. We assume next that the time and parameter dimensions are both discrete, which means that any speech coefficient is represented by a couple of positive integers. M is thus defined as:

$$M : (\mathbb{N}, \mathbb{N}, \Omega) \rightarrow D$$

M further depends on a domain Ω that contains all the other information required to compute the mask.

Two solutions are commonly used in the missing data community for the mask domain D:

- $D = \{0, 1\}$: Such masks are usually called *hard* masks: any coefficient is either masked or not.
- $D = [0, 1]$: Such masks are usually called *soft* masks. Their value is interpreted as the probability that the coefficient is masked.

Several missing data studies have shown that soft masks give better results than hard masks, at least for the classical missing data recognition systems.

Let us now consider the "decision" domain Ω. A coefficient is usually masked when, for this coefficient, the contribution of the useful signal is dominated by the contribution of the noise. Such a criterion can be interpreted as applying a threshold to the local signal-to-noise ratio (SNR). Sometimes Ω is not limited to the local SNR, but has additional constraints, such as minimum duration and bandwidth for masked fragments, or smoothness of the masked values over time and frequency.

The SNR-based definition of Ω is the most intuitive, and the most widely accepted in the speech recognition community, but alternative definitions can be found, depending on the speech-processing task. When the final objective is to denoise the signal in order to increase its SNR, then this definition is quite appropriate. But real applications often aim instead at improving the intelligibility of the speech signal, in which case an increase in SNR does not systematically translate into an increase in speech intelligibility. This is also true when the objective is to recognize the speech signal. In such cases, it might be better to base the decision on word accuracy rather than on local SNR. In the next section, we present some experimental results and discussion on this point.

3.2 WER-based oracle masks

For SNR-based masks, the decision domain Ω is mainly composed of the local SNR. However, for masks based on word accuracy, Ω may also include every piece of information and all processes that influence the final word recognition, i.e., the whole speech recognition system.

To compute such masks, a simplistic technique would be to test every possible mask on each sentence and to compare the recognition accuracies. This is obviously not feasible, because of the combinatory problem. We thus have decided to approximate this optimal solution with the following heuristic:

- First, the speech models are force-aligned on the clean signal to obtain the "best possible" alignment between the models and the signal. Let us call $e_0(t)$ the resulting model state aligned with frame $y(t)$.

- Then, the same speech models are aligned on the noisy signal, with the classical Viterbi algorithm. This gives the "baseline" alignment. Let us call $e(t)$ the resulting model state aligned with frame $y(t)$.
- For every feature dimension i, its contribution to the baseline alignment is defined as:

  $$p(y_i(t)|e(t)) = \int \cdots \int p(y(t)|e(t))dy_1(t) \cdots dy_{k \neq i}(t)$$

  In this contribution, the likelihood is marginalized over all coefficients except the i^{th} one.
- Similarly, the contribution of coefficient i to the best possible alignment is:

  $$p(y_i(t)|e_0(t)) = \int \cdots \int p(y(t)|e_0(t))dy_1(t) \cdots dy_{k \neq i}(t)$$
- Then, the difference between both contributions is computed and compared to 0 to decide whether the i^{th} coefficient is masked or not:

  i is masked if and only if $p(y_i|e) - p(y_i|e_0) > 0$.

  Intuitively, the coefficients that maximise this difference strongly influence the recognizer to prefer the baseline alignment over the best possible one. Consequently, they are masked.

These WER-based oracle masks are compared in figures 1 and 2 with classical SNR-based oracle masks on the standard Aurora2 noisy corpus. Aurora2 is composed of sequences of digits with different kinds of additive noise at different signal-to-noise ratio (SNR). Aurora2 is a standard database, and the reference results reported in figures 1 and 2 are respectively taken from [7] and [13]. In both experiments, I have used the HTK toolkit [19], modified to perform hard and full (unbounded) missing data marginalization on cepstral features.

Note that it is impossible to conclude from this experiment that WER-based masks are usable in practice, because it is realized with oracle masks, which are not available in a real situation. The objective is thus only to prove that the approximations described above to compute the coefficients likelihood contribution and to derive oracle masks are valid. These results, which undoubtly show the potential of WER-based masks, also prove that marginalization with such masks produce good enough results, so that it is worthwhile investigating methods to model or estimate these masks, such as the one described in the next section.

3.3 Confidence-based masks

We have shown in the previous section that WER-based oracle masks can be an efficient alternative to traditional SNR-based masks. However, we still have to prove that these masks can be computed from the information available at test time.

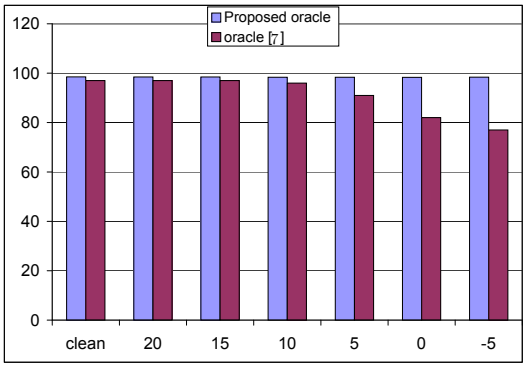

Figure 1: Comparison of the proposed WER-based oracle mask with the SNR-based oracle mask from [7]

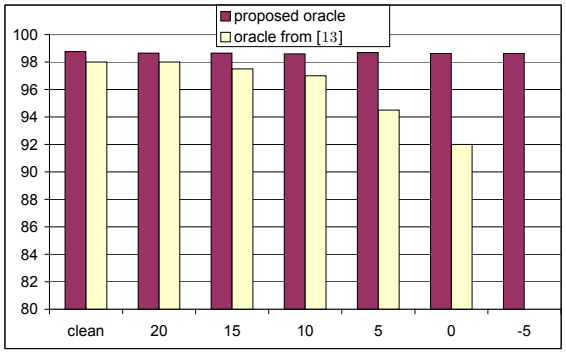

Figure 2: Comparison of the proposed WER-based oracle mask with the SNR-based oracle mask from [13]

A few years ago, when missing data recognition was only at its very beginning, researchers have faced the very same challenge, but for SNR-based masks. The first solutions proposed were based on signal processing algorithms, such as spectral subtraction, which estimates the local SNR in every frequency band.

However, SNR-information, from which SNR-based masks were successfully derived, might not be relevant any more with WER-based masks. We rather propose to exploit speech recognition confidence measures, which can be viewed as WER estimators.

For this purpose, we have chosen the state-of-the-art speech recognition confidence measure of Wessel et al. [18]. This measure computes a posteriori word probabilities using a forward-backward algorithm on the speech recognition trellis.

Once a confidence value is computed for every recognized word, the algorithm proceeds by thresholding this value in order to select only the words that are the most likely to be erroneous. Masks are then built for these "bad" words only, while all the other words that

are above the confidence threshold are assumed to be correct and are not masked.

The procedure that has been followed to build these masks is largely inspired by the algorithm chosen to compute oracle masks in section 3.2. Assuming that the word that has been recognized is wrong, our objective is to mask the acoustic observations that tend to support this word, or in other words, that contribute the most to the likelihood of this word.

Let $[y(t_1), \cdots, y(t_2)]$ be the feature frames aligned with one presumably wrong word. Each feature vector is composed of N dimensions:

$$y(t) = [y_1(t), \cdots, y_i(t), \cdots, y_N(t)]^t$$

All coefficients $(y_i(t))_{t_1 \leq t \leq t_2, 1 \leq i \leq N}$ are pooled together, irrespectively of their time indexes, and are sorted according to their contribution $p(y_i(t)|e(t))$ to the word likelihood, as defined above. Finally, the M coefficients with the largest contribution are masked, where M is a fixed predefined mask density constant.

3.4 Discussion: limits of the approach

One may question the fact that the proposed approach belongs to the missing data speech recognition research area. Indeed, the fundamental principle of missing data recognition consists in segregating speech from noise, whereas in this work, WER-based masks do not necessarily discriminate between clean and noisy coefficients any more. However, WER-based masks can be seen as a generalization of SNR-based masks, because they do not only mask coefficients that are likely to be corrupted by additive noise, but also coefficients that encode uncommon/unreliable acoustic information that has a negative impact on speech recognition accuracy.

Alternatively, the proposed method can also be viewed as an original solution to include confidence measures within speech recognizers. Indeed, computing confidence measures for speech recognition is a very important and challenging topic that is still the focus of many researches. One of the major application issue in this domain is how to efficiently exploit the confidence measures that have been computed to improve speech recognition accuracy. The most common solution consists in reweighting the graph, trellis or n-best solutions at the output of the speech recognizer by the confidence score, so that paths with a lower initial score, but composed of words that have a higher confidence, are pulled up. However, the main drawback of this solution is that all competing word sequences are derived from the very same acoustic information: when this information is erroneous, then there is no reason why any of these competing solutions should be better than the first one !

Conversely, the solution devised in this work alters the acoustic information based on the confidence measure, so that the second run of the recognizer uses only an (hopefully) better subset of the original acoustic information: this intuitively explains why the second result should be better than the first one.

This reasoning also points out a weakness of the current method: indeed, a word can have a low confidence score because of a deficient speech recognition language model, and not because of acoustics. This is typically the case for two words that have the same pronunciation but different graphemes, such as *air* and *heir*. Then, masking acoustic information will not help to correct such errors. However, the impact on the results of such a case may not be very dramatic, thanks to at least two reasons:

- Assuming the confidence measure is correct, the word is in any case erroneous, and there is only a little chance that this process introduces additional errors;
- If the recognized phones sequence is correct, it is still likely that masking some acoustic information will not modify it. Actually, it is a well-known fact in the missing data recognition community that a large fraction of spectral coefficients can be masked without affecting recognition accuracy, when there is no noise and if the masks are chosen so that a single large contiguous frequency region is not masked.

Therefore, the only case where the WER might increase occurs when the confidence measure is wrong.

4 Experimental validation

4.1 Experimental set-up

A brief experimental validation of this approach is realized on the ESTER corpus, which is a speech database composed of broadcast news recordings from french radios [10]. This corpus has been used in the french Technolangue broadcast news transcription evaluation that has been conducted in 2004.

We have used the ANTS transcription system developed in our team as a baseline speech recognizer [3]. Its main properties are:

- First, the audio stream is partitioned into telephone *versus* broadband audio segments;
- Then, another segmentation into male/female speakers is realized based on dedicated Gaussian Mixture Models;
- The remaining segments are parameterized with the HTK toolkit into 39 MFCC parameters with first and second order derivatives and with batch cepstral mean normalization;

- Gender-independent triphone acoustic models, which have been previously adapted with MAP adaptation on respectively male and female speech, are used to recognize every segment.

  Recognition is performed with the Julius decoder [12], which operates in two passes: a first left-to-right pass with approximate right triphone context and bigram language model generates a recognition trellis, and a second A* reverse pass with words trigrams outputs the final result. A predefined 60000 words-lexicon is used.

This baseline system has been modified as follows to implement the approach described in this work:

- A local implementation of the Wessel et al. confidence measure realizes a forward-backward pass on the trellis computed in the first decoding pass to estimate words log-posterior probabilities;
- The words that have been recognized in the first pass and that have a log-confidence measure lower than -0.5 are passed to the masking module, which computes the likelihood contributions of all the coefficients belonging to a given word segment, sorts them and masks the ones with the highest contributions so that the mask density is 3 %.
- The julius decoder has been enhanced with hard missing data full-marginalization; a second run of the decoder is realized with this modified julius and the masks that have been computed.

Both constants used in this procedure have been heuristically defined prior to validation.

4.2 Broadcast news results

Table 1 reports word error rates (WER) on one hour of the ESTER test corpus for both the baseline transcription and proposed system. The confidence interval is $\pm$ 0.5 %.

Recognition system	WER [%]
Baseline	18.6
Confidence-based mask	17.5

Table 1: Experimental results of the baseline system and proposed approach

5 Conclusion and future work

We have proposed in this work a new masking paradigm for robust speech recognition that is based on WER optimization and that exploits confidence measures. This approach has several theoretical advantages compared to classical missing data SNR-based masks. First, it directly optimizes the word error rate, which is also the final objective criterion of speech recognition. Second, it does not make any difference between static and dynamic parameters: they are all treated equally, which solves a classical issue of missing data recognition systems, where different heuristics have been proposed to mask dynamic coefficients. Third, the proposed approach is independent of the front-end, which means that it can directly be used with any new parameterization approach, and it can be combined with every denoising methods.

Preliminary experiments reported have shown the potential of these WER-based masks, and also that it is possible to estimate them from confidence measures. However, the approach sketched in this work can be improved in many aspects:

- Confidence information should be better exploited than with a simple threshold, for example by training models, such as HMMs, neural networks, etc.
- Hard decisions about which segments are masked are probably not the best possible option. It shall be better to use soft decisions or weights to mask segments.
- For now, only confidence information is used, but intuitively, other kinds of features shall be helpful. For instance, global SNR might give a good indication about the expected mask density.
- The mask density for erroneous segments is fixed a priori: it shall be better to give it a parametric form in function of the value of the confidence measure or of the segmental SNR for instance.
- For now, all the frames in a word segment have the same probability of being masked, but it shall be beneficial to rather estimate the confidence measure for each phone within a word, for example by comparing the pronunciations of competing words, in order to avoid masking correctly recognized phones.
- Soft masks have proven to be better than hard masks, and soft WER-masks should be considered as well.

References

[1] A. Adami, L. Burget, S. Dupont, H. Garudadri, F. Grezl, H. Hermansky, P. Jain, S. Kajarekar, N. Morgan, and S. Sivadads. QUALCOMM-ICSI-OGI for ASR. In *Proc. ICSLP*, 2002.

[2] J. Baker, M. Cooke, and D. P. W. Ellis. Decoding speech in the presence of other sources. *Speech Communication*, 45(1):5–25, January 2005.

[3] A. Brun and al. Ants le système de transcription automatique du loria. In *Séminaire ESTER, Avignon*, Mar 2005.

[4] C. Cerisara. Towards missing data recognition with cepstral features. In *Proc. EUROSPEECH'2003*, Geneva, Switzerland, September 2003.

[5] C. Cerisara, S. Demange, and J.-P. Haton. On noise masking for automatic missing data speech recognition: a survey and discussion. *Computer Speech and Language*, 21(3):443–457, July 2007.

[6] M. Cooke. A computer model of peripheral auditory processing incorporating phase-locking, suppression, and adaptation effects. *Speech Communication*, 1986.

[7] M. Cooke, P. Green, L. Josifovski, and A. Vizinho. Robust automatic speech recognition with missing and unreliable acoustic data. *Speech Communication*, 34(3):267 – 285, June 2001.

[8] L. Deng, J. Wu, J. Droppo, and A. Acero. Analysis and comparison of two speech feature extraction/compensation algorithms. *IEEE Signal Processing Letters*, 12(6):477–480, June 2005.

[9] Speech processing, transmission and quality aspects (stq); distributed speech recognition; advanced front-end featur extraction algorithm; compression algorithms. http://www.etsi.org, 2003.

[10] G. Gravier, J.-F. Bonastre, S. Galliano, E. Geoffrois, K. Mc Tait, and K. Choukri. Ester, une campagne d'évaluation des systčmes d'indexation d'émissions radiophoniques. In *Proc. JEP*, Fez, 2004.

[11] J. Häkkinen and H. Haverinen. On the use of missing feature theory with cepstral features. In *CRAC Workshop*, Aalborg, Danemark, September 2001.

[12] A. Lee, T. Kawahara, and K. Shikano. Julius - an open source real-time large vocabulary recognition engine. pages 1691–1694, 2001.

[13] A.C. Morris, J. Barker, and H. Bourlard. From missing data to maybe useful data: Soft data modelling for noise robust ASR. In *Proc. WISP*, number 06, Stratford-upon-Avon, England, April 2-3 2001.

[14] Philippe Renevey. *Speech recognition in noisy conditions using missing feature approach*. PhD thesis, Ecole Polytechnique Fédérale de Lausanne, 2001.

[15] S. Srinivasan and D.L. Wang. Transforming binary undertainties for robust speech recognition. *IEEE Trans. on Audio, Speech and Language Processing*, 2007.

[16] H. Van Hamme. Robust speech recognition using missing feature theory in the cepstral or lda domain. In *Proc. EUROSPEECH'2003*, pages 3089–3092, Geneva, September 2003.

[17] H. Van Hamme. PROSPECT features and their application to missing data techniques for robust speech recognition. In *Proc. ICSLP*, volume 1, pages 101–104, Jeju Island, Korea, October 2004.

[18] F. Wessel, R. Schlüter, K. Macherey, and H. Ney. Confidence measures for large vocabulary continuous speech recognition. *IEEE Trans. SAP*, 9:288–298, 2001.

[19] P. C. Woodland and S. J. Young. The htk continuous speech recogniser. pages 2207–2219, Berlin, 1993.