Accurate marginalization range for missing data recognition

Sébastien Demange, Christophe Cerisara, Jean-Paul Haton

LORIA-UMR 7503
54500 Vandoeuvre-les-Nancy - FRANCE
sebastien.demange, christophe.cerisara,jean-paul.haton@loria.fr

Abstract

Missing data recognition has been proposed to increase noise robustness of automatic speech recognition. This strategy relies on the use of a spectrographic mask that gives information about the true clean speech energy of a corrupted signal. This information is then used to refine the data process during the decoding step. We propose in this work a new mask that provides more information about the clean speech contribution than classical masks based on a Signal to Noise Ratio (SNR) thresholding. The proposed mask is described and compared to another missing data approach based on SNR thresholding. Experimental results show a significant word error rate reduction induced by the proposed approach. Moreover, the proposed mask outperforms the ETSI advanced front-end on the HIWIRE corpus.

Index Terms: robust speech recognition, missing data, bounded marginalization

1. Introduction

The presence of background noise typically causes mismatches between training and testing conditions, which significantly degrade the performance of automatic speech recognizers (ASRs). Over the last decades, many solutions to reduce the effect of noise have been proposed. Acoustic models can be adapted to new noisy conditions, the analysis front-end can be made robust to noise, and noise reduction algorithms can be used as preprocessing stages.

Although many of these methods have shown superior performance in noisy conditions compared to standard speech recognition, noise robustness is still a challenging issue for nowadays speech recognizers, especially for non-sationary noise.

More recently, speech recognition with missing data has been proposed. This technique relies on a clustering of spectral features into two classes: time-frequency (T-F) units of a noisy speech signal that contain more speech energy than noise energy are classified as reliable data, while T-F units containing more noise energy are classified as missing data. Hence, the resulting clustering produces a binary mask that is exploited in missing data recognition techniques [1].

Oracle masks are computed from the exact local signal-to-noise ratio (SNR). Recognition results with oracle masks can be interpreted as the upper limit of missing data recognizers. The very good results usually reported with oracle masks show the potential of missing data recognition.

There are two common techniques to handle missing data masks during recognition. The first one involves replacing all missing coefficients by their estimated value to provide the recognizer a completely observed signal. This approach is called *data-imputation* [2]. The second one, called *data-marginalization*, relies on a modified recognizer engine designed to handle partial feature vectors. The observation likelihood is thus integrated over all possible values of the unknown clean speech. Several data marginalization schemes have been proposed depending on the available information about the true value of missing features [3, 4]. This information is usually contained in the missing data mask itself, as discussed in details in [3].

We propose in this paper a new type of missing data mask, which is designed to provide a more accurate information about the true value of clean speech than classical masks based on a local SNR threshold.

The organization of the paper is as follows. In section 2, the background and motivation that led to this mask definition is presented. The proposed approach is then detailed in section 3, while experimental results are given in sections 4 and 5.

2. Motivations

Assuming diagonal covariances, and that observations can be missing or uncertain, the likelihood $p(Y_i|\Theta, Q)$ of a particular noisy spectral data Y_i should be replaced by its expected value $\hat{p}(Y_i|\Theta, Q)$ [3]:

$$\hat{p}(Y_i|\Theta, Q) = \int_{\mathcal{D}} p(x_i|\Theta, Q).p(x_i|Y_i, m_i)dx_i \quad (1)$$

where Θ is the set of speech models parameters, Q is the state aligned with Y, and $p(x_i|Y_i, m_i)$ is the evidence pdf of the clean speech energy conditionned on the mask m_i and on the noisy observation Y_i and defined on the domain $\mathcal{D}$.

For example, any static spectral coefficient is usually considered as missing when its local SNR is below 0 dB, and reliable otherwise. Let $\gamma(\cdot)$ be the function (classically a logarithm or a cubic root) used to compress the spectral power. According to noise additivity in the spectral domain, the speech contribution X_i of any masked observation coefficient Y_i lies in the interval:

$$\gamma(0) \leq X_i \leq \gamma(\gamma^{-1}(Y_i)/2) \quad (2)$$

Conversely, the speech contribution X_i of any observed reliable coefficient Y_i lies in the interval:

$$\gamma(\gamma^{-1}(Y_i)/2) \leq X_i \leq Y_i \quad (3)$$

Let $Y_{i,snr0} = \gamma(\gamma^{-1}(Y_i)/2)$. Each inequality (2) and (3) defines respectively a uniform evidence pdf $u(0, Y_{i,snr0})$ and $u(Y_{i,snr0}, Y_i)$ for missing and reliable data. Hence, equation (1) can be rewritten as:

$$\begin{aligned}\hat{p}(Y_i|\Theta,Q) &= m_i.\int_0^{Y_{i,snr0}} p(x_i|\Theta,Q).u(0,Y_{i,snr0})dx_i\\ &\quad + (1-m_i).\int_{Y_{i,snr0}}^{Y_i} p(x_i|\Theta,Q).u(Y_{i,snr0},Y_i)dx_i\\ &= \frac{m_i}{Y_{i,snr0}}.\int_0^{Y_{i,snr0}} p(x_i|\Theta,Q)\,dx_i\\ &\quad + \frac{1-m_i}{Y_i\ -\ Y_{i,snr0}}.\int_{Y_{i,snr0}}^{Y_i} p(x_i|\Theta,Q)\,dx_i \qquad (4)\end{aligned}$$

This approach is used for example in [5]. In the following, this marginalization method is called **SNR-0 mask**, because the marginalization interval is derived from a threshold that is usually equal to a local SNR of 0 dB.

Equation (4) shows that the marginal output likelihoods depend on the marginalization range, which thus impacts the final recognition performance.

The least informative case occurs with infinite marginalization ranges, in which case the model output likelihoods doe not depend on the observation:

$$\hat{p}(Y_i|\Theta,Q) \quad = \quad \int_0^{\infty} p(x_i|\Theta,Q).p(x_i|Y_i,m_i)dx_i \quad (5)$$

In contrast, when the marginalization interval contains X_i and when its bounds tend to get as close as possible to X_i, the missing data output likelihood becomes equal to the ideal likelihood computed with a standard recognizer on clean speech:

$$\begin{aligned}\hat{p}(Y_i|\Theta,Q) &= \lim_{\xi\to 0}\int_{X_i-\xi}^{X_i+\xi} p(x_i|\Theta,Q).p(x_i|Y_i,m_i)dx_i\\ &= p(X_i|\Theta,Q) \qquad (6)\end{aligned}$$

The objective of this work is thus to propose a method to reduce the marginalization interval around the true value of the clean speech contribution, in order to improve the accuracy of the missing data model likelihoods.

3. Proposed mask

In this section, a method is described to build a missing data mask that provides more accurate bounds for the true clean speech value X_i of an observed noisy value Y_i than traditional SNR-based masks. First, the ratio $\frac{X_i}{Y_i}$ between clean and noisy speech is computed on a stereo training corpus. This results in a time-frequency representation that provides for every noisy spectral feature the relative contribution of the clean speech energy. This ratio is related to the local SNR as follows:

$$\frac{X_i}{Y_i} \quad = \quad \frac{1}{1+10^{-\frac{\mathrm{SNR_{local}}}{20}}} \qquad (7)$$

In a second step, the set of all training $\frac{X_i}{Y_i}$ is clustered into K classes $(M^k)_{k\in[1,K]}$. Each cluster M^k is represented by a mean vector $\mu^k = (\mu_1^k,\mu_2^k,\ldots,\mu_i^k,\ldots,\mu_N^k)^T$ and a diagonal covariance matrix $\Sigma^k = \mathrm{diag}(\sigma_1^k,\sigma_2^k,\ldots,\sigma_i^k,\ldots,\sigma_N^k)$ where N is the size of static spectral vectors.

These K clusters M^k represent our set of potential missing data masks, where each mask is associated to accurate marginalization intervals.

Let $Y = (Y_1,\ldots,Y_N)^T$ be a noisy observation vector of the test corpus and $X = (X_1,\ldots,X_N)^T$ the (unknown) clean speech contribution in Y. Let M^k be the missing data mask associated to Y.

By definition,

$$p(\beta \le \frac{X_i}{Y_i} \le \gamma|M^k) = \int_\beta^\gamma \mathcal{N}(x_i;\mu_i^k,\sigma_i^k)dx_i \qquad (8)$$

We can note that:

$$p(\mu_i^k - 2.\sigma_i^k \le \frac{X_i}{Y_i} \le \mu_i^k + 2.\sigma_i^k|M^k) = 0.95 \qquad (9)$$

Values outside this interval are neglected since they represent only 5 % of the possible values of $\frac{X_i}{Y_i}$. Then we can assume:

$$\mu_i^k - 2.\sigma_i^k \le \frac{X_i}{Y_i} \le \mu_i^k + 2.\sigma_i^k \qquad (10)$$

So, for a given noisy observation and its corresponding mask M^k, the clean speech energy X_i is bounded as follow:

$$Y_i.(\mu_i^k - 2.\sigma_i^k) \le X_i \le Y_i.(\mu_i^k + 2.\sigma_i^k) \qquad (11)$$

These bounds will be used in the following experiments. Finally, the marginal output likelihood is:

$$\hat{p}(Y_i|\Theta,Q) \quad = \quad \int_{Y_i.(\mu_i^k-2.\sigma_i^k)}^{Y_i.(\mu_i^k+2.\sigma_i^k)} p(x_i|\Theta,Q).p(x_i|Y_i,m_i)dx_i$$

With a reasonable number of clusters, cluster variances are expected to be small enough to produce small marginalization intervals and consequently to reduce the confusion between acoustic models during decoding.

4. Experimental setup

4.1. Database

The proposed method has been evaluated on the HIWIRE database. This database has been collected and packaged under the auspices of the IST-EU STREP project HIWIRE [6]. The database contains 8 100 English utterances pronounced by 81 non-native speakers (31 French, 20 Greek, 20 Italian and 10 Spanish speakers). The collected utterances correspond to human input in a command and control aeronautics application defining a close set of 133 words. A finite state grammar is used for the recognition tasks. The data has been recorded in studio with a high quality close-talking microphone at 16 kHz. Real noise recorded in an airplane cockpit was artificially added to the data. The signals are provided in clean, low (LN), mid (MN) and high (HN) noise conditions. The three noise levels correspond approximately to SNRs of 10 dB, 5dB and -5dB respectively. The 50 first utterances of each speaker are used as training and development corpus, while the 50 remaining ones are used for testing.

4.2. Acoustic models

Since the HIWIRE database does not provide sufficient training material, acoustic models are pre-trained on the AURORA4 [7] clean training set. The resulting models are further reestimated on the HIWIRE training corpus with 10 Baum-Welch iterations. The models are context free, 3-states HMM phone models. Each state contains a mixture of 128 Gaussians. The feature domain, which is also the marginalization domain of missing data, is the 12-bands Mel spectral domain with cube-root compression

of the speech power. Temporal derivatives are further added, leading to a 24 dimensional feature vector.
Dynamic features are computed from their static counter-parts. The following equation is used for this purpose:

$$\Delta Y_i(t) = \frac{\sum_{j=-N}^{j=N} j.Y_i(t+j)}{\sum_{j=-N}^{j=N} j^2} \tag{12}$$

Instead of marginalizing them, we propose to estimate their clean value $\Delta\hat{X}_i(t)$ as follows:

$$\Delta\hat{X}_i(t) = \frac{\sum_{j=-N}^{j=N} j.\hat{X}_i(t+j)}{\sum_{j=-N}^{j=N} j^2} \tag{13}$$

where $\hat{X}_i(t+j)$ is infered from the noisy observation $Y_i(t+j)$ and its mask M^k:

$$\hat{X}_i(t+j) = \arg\max_x \; p(X_i(t+j) = x | M^k) \tag{14}$$

$$\hat{X}_i(t+j) = Y_i(t+j).\mu_i^k \tag{15}$$

4.3. Missing data masks

Oracle mask: Oracle masks are derived from the perfect knowledge of the contribution of speech energies in noisy signals. Let $R^t = (\frac{X_1^t}{Y_1^t}, \ldots, \frac{X_i^t}{Y_i^t}, \ldots, \frac{X_N^t}{Y_N^t})^T$ be the clean to noisy speech ratio at time t. The oracle mask at time t is then the mask $M^{\tilde{k}}$ with

$$\tilde{k} = \arg\max_k \; \mathcal{N}(R^t; \mu^k, \sigma^k) \tag{16}$$

Estimated mask: In practice, only noisy speech is observed. In the following experiments, $K = 64$ masks have been clustered. For each mask M^k with $k \in [1, 64]$, a Gaussian mixture model (GMM) with 64 Gaussians is trained on the noisy frames aligned with M^k from the HIWIRE training set. An ergodic HMM with 64 states is then built where each state respectively contains one of the k masks. The HMM transition probabilities are trained from the sequences of masks observed on the training corpus. Estimated masks correspond to the HMM state sequence given by Viterbi decoding. Finally, marginalization bounds $[b_i(Y_i), b_s(Y_i)]$ are infered from masks as described in section 3. We consider in the following experiments that $p(x_i) = C^{te}, x_i \in [b_i(Y_i), b_s(Y_i)]$.

5. Results

A new measure is introduced in section 5.1 to evalute the capacity of the proposed masks to reduce the marginalization range. The masks are then evaluated in terms of speech recognition accuracy in section 5.2, and compared with the SNR-0 masks (see equation 1). The proposed method is also compared with the noise-robust standard ETSI AFE parameterization. WER are given with a confidence interval of about ± 0.5 %.

5.1. Marginal MSE based evaluation

We introduce next the Marginal Mean Square Error (MaMSE). MaMSE is a measure that represents how well the marginalization intervals encapsulate the true values X_i of the speech energies. MaMSE is defined as follows for a single speech frame:

$$MaMSE = \frac{1}{N} \sum_{i=0}^{i=N} \int_{b_i(Y_i)}^{b_s(Y_i)} (X_i - x_i)^2 \, p(x_i | Y_i, m_i) dx_i$$

where N is the number of coefficients.

			SNR-0 mask		Proposed mask	
			IN	OUT	IN	OUT
Oracle masks	Clean	MaMSE	10	-	0.02	-
		%	100	0	100	0
	LN	MaMSE	63	202	10	37
		%	99	1	97	3
	MN	MaMSE	99	295	14	53
		%	99	1	97	3
	HN	MaMSE	266	618	26	94
		%	99	1	95	5
Estimated masks	Clean	MaMSE	41	536	2	97
		%	99	1	99	1
	LN	MaMSE	63	62	12	44
		%	94	6	81	19
	MN	MaMSE	99	306	16	70
		%	93	7	77	23
	HN	MaMSE	272	590	29	190
		%	95	5	73	27

Table 1: *Comparison of MaMSE for the SNR-0 (see equation 1) and proposed masks. The true clean speech contribution may be either in the marginalization interval (IN) or outside (OUT). The relative proportion of IN and OUT coefficients are given along with the value of MaMSE.*

Table 1 presents comparative MaMSE measures for the proposed and SNR-0 masks. These measures are computed on the whole HIWIRE test set. The MaMSE score is always better for the proposed mask than for the SNR-0 mask. This confirms that the marginalization intervals are indeed smaller in the new approach. On the other hand, the true clean speech energy is more often out of the marginalization interval for the proposed mask than for the SNR-0 mask. For instance, in high noise conditions with estimated mask, the proposed approach gives 27 % of mask error while the competing approach gives only 5 %. The best compromise between the width of the marginalization interval and the mask errors is assessed next in terms of speech recognition accuracy.

5.2. WER based evaluation

Table 2 shows for each noise condition the recognition word error rate (WER) obtained on the HIWIRE test set. The results obtained with the proposed mask always outperform those obtained with the SNR-0 mask.

Figures 1 and 2 graphically compare the performance of the proposed and SNR-0 masks, respectively with oracle and estimated masks. The recognition results obtained with the proposed oracle masks (figure 1) clearly show the good potential of the proposed approach with a maximum WER of 19.7% for HN condition. However, with estimated masks (figure 2), the performance of the proposed approach is not as spectacular than with oracle masks: it is still better than SNR-0 masks, but only of 1 % in HN condition. As mentionned in the previous section, this loss of performance is due to mask errors that are more frequent in such drastic conditions (see table 1). Furthermore, the impact of such errors is worse with small marginalization intervals than with larger ones. Nevertheless, the proposed approach

		Clean	LN	MN	HN
ETSI AFE		3.6	16.2	34.8	93.0
SNR-0 mask	Oracle	5.6	25.6	32.6	67.4
	Estimated	18.0	41.9	41.8	73.8
Proposed Mask	Oracle	10.8	9.8	12.1	19.7
	Estimated	10.8	14.8	24.4	72.1

Table 2: *Recognition word error rates on the HIWIRE test corpus for respectively the proposed and SNR-0 masks. The results with a standard recognizer based on the ETSI AFE are also given for comparison.*

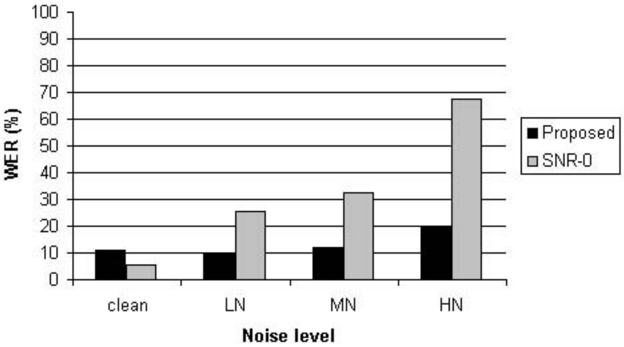

Figure 1: *Performance comparison between the proposed and SNR-0 oracle masks.*

is largely superior to SNR-0 masks on low and medium noise levels.

Figure 3 graphically compares the proposed approach with the standard noise robust method ETSI AFE. The training and test procedure of the ETSI AFE acoustic models is exactly the same as for the proposed approach (see section 4.2).

Apart from clean speech, in which case it is anyway better not to marginalize at all, the proposed approach gives the best recognition accuracies.

6. Conclusion

We have proposed in this work a new missing data recognition approach, in which reduced marginalization intervals are computed for each possible mask. The set of all possible masks and intervals is obtained by clustering on a clean and noisy stereo training corpus. The main principle of the proposed approach consists in training accurate marginalization intervals that are as small as possible, in order to improve the precision of marginalization. This results in a new compromise between the width of these intervals and the masking errors, which occur when the true speech energy is outside the interval.

The proposed approach has been evaluated on the HIWIRE corpus, and experimental results show the effectiveness of the proposed mask compared to classical missing data masks based on a SNR threshold. A significant WER reduction is also observed when compared with the standard robust front-end ETSI AFE, which is known to be much more robust than classical cepstral parameters.

7. 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.

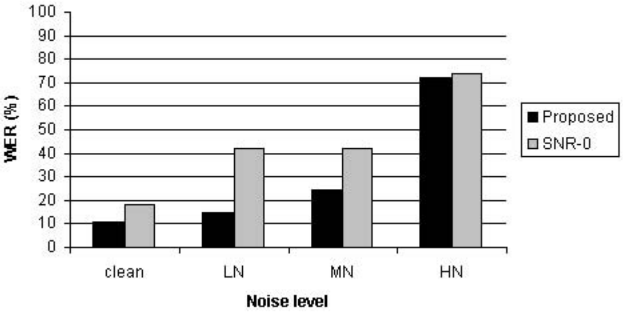

Figure 2: *Performance comparison between the proposed and SNR-0 estimated masks.*

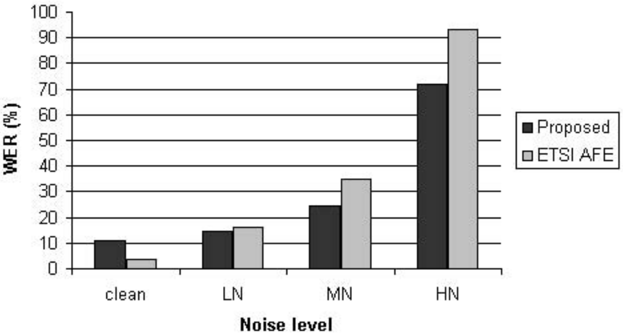

Figure 3: *Performance comparison between the proposed estimated mask and the ETSI AFE acoustic models.*

8. References

[1] 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*, vol. 21, no. 3, pp. 443–457, July 2007.

[2] B. Raj, *Reconstruction of incomplete spectrograms for robust speech recognition*, Ph.D. thesis, Carnegie Mellon University, 2000.

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

[4] 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.

[5] S. Demange, C. Cerisara, and J-P. Haton, "Missing data mask models with global frequency and temporal constraints," in *Proc. ICSLP*, Pittsburgh,Pennsylvania/USA, September 2006.

[6] A. Potamianos, G. Bouselmi, D. Dimitriadis D. Fohr, R. Gemello, F. Illina, P. Maragos, M. Matassoni, V. Pitsikalis, J. Ramirez, E. Sanchez-Soto, J.C. Segura, and P. Swaizer, "Towards speaker and environmental robustness in asr: The hiwire project," in *Proc, Workshop on Speech Recognition and intrinsic Variation*, Toulouse,France, May 2006.

[7] N. Parihar and J. Picone, "Analysis of the aurora large vocabulary evaluations," in *Proc. EUROSPEECH*, Geneva, Switzerland, September 2003, vol. 4, pp. 337–340.