Exploiting models intrinsic robustness for noisy speech recognition

Christophe Cerisara, Dominique Fohr, Odile Mella, Irina Illina

Exploiting models intrinsic robustness for noisy speech recognition

Christophe Cerisara, Dominique Fohr, Odile Mella,
Irina Illina

Speech Group, LORIA UMR 7503
BP 239 - 54506 Vandoeuvre
FRANCE
http://www.loria.fr/equipes/parole

Abstract

We propose in this paper an original approach to build masks in the framework of missing data recognition. The proposed soft masks are estimated from the models themselves, and not from the test signal as it is usually the case. They represent the intrinsic robustness of model's log-spectral coefficients. The method is validated with cepstral models, on two synthetic and two real-life noises, at different signal-to-noise ratios. We further discuss how such masks can be combined with other signal-based masks and noise compensation techniques.

1. Introduction

A number of previous studies have shown that the combination of speech and noise is well approximated in the log-spectral domain by the masking paradigm [1]: At any time, every single frequency coefficient is dominated either by speech or noise. This observation lead to a number of methods that aim at separating the speech from the noise in the log-spectrum.

One such method is the multi-band approach [2] [3], where the frequency domain is divided into several bands, and one recognizer is trained within each band. When recombining the sub-band recognition results, the local Signal-to-Noise Ratio (SNR) can be used to weight each sub-band.

The missing data recognition approach rather considers a full-band recognizer with a time-frequency mask that annotates every coefficient with its emitting source, either speech or noise. The mask can be taken into account either by marginalizing the noisy coefficients during likelihood computation [4], or by "denoising" the masked coefficients before recognition [5]. In the latter case, a model of the clean speech given the noisy observation is trained. Note that a similar algorithm has been proposed in [6], where it is applied for denoising in the whole frequency range.

The most difficult part of missing data recognition approaches is to build a mask that represents as accurately as possible the noisy fragment in the spectro-temporal domain. The most successful methods proposed so far to estimate such masks are described in:

- [7], where a mask model is trained on clean speech databases corrupted by artificial noise;
- [8], where potential masks are first estimated using signal processing techniques and are then chosen during decoding by maximizing the final likelihood over all the possible masks.

Usually, the masks are built in order to filter out noisy coefficients. We propose in this work an original point of view, where our "masks" represent the models coefficients that are the most robust to noise. Therefore, the masks are not estimated any more on the test signal, but are rather computed once at training time and are then considered as constants.

2. Definition of coefficients robustness

The *robustness* is defined in the log-spectrum, as the diagonal matrix R whose coefficients are related to the average energy of the Gaussian. We assume that the most energetic frequencies of speech have the highest local SNR and are thus the most robust to noise.

R is applied in the log-spectrum by multiplying each coefficient of both observation X^l and Gaussian mean μ^l by R. The log-likelihood thus becomes:

$$\begin{aligned} & \log P(X^l|\mu^l,\Sigma^l) \\ = \; & K - \frac{1}{2}(R(\mu^l - X^l))^T {\Sigma^l}^{-1} R(\mu^l - X^l) \\ = \; & K - \frac{1}{2}(\mu^l - X^l)^T R {\Sigma^l}^{-1} R(\mu^l - X^l) \end{aligned} \quad (1)$$

Applying the robustness is thus equivalent to replacing the covariance matrix by $R^{-1} \cdot \Sigma^l \cdot R^{-1}$.

Let us now consider the cepstral model $N(\mu^c, \Sigma^c)$.

Computing the log-likelihood gives:

$$\begin{aligned} & \log P(X^c|N(\mu^c,\Sigma^c)) \\ = \; & K - \frac{1}{2}(\mu^c - X^c)^T \Sigma^{c-1}(\mu^c - X^c) \\ = \; & K - \frac{1}{2}(\mu^l - X^l)^T D^T \Sigma^{c-1} D(\mu^l - X^l) \quad (2) \end{aligned}$$

where D is the DCT matrix. The vector difference in the log-spectral domain can now be weighted by the robustness:

$$\begin{aligned} & \log P(X^c|N(\mu^c,\Sigma^c)) = K - \\ & \frac{1}{2}(\mu^l - X^l)^T R D^T \Sigma^{c-1} D R(\mu^l - X^l) \quad (3) \end{aligned}$$

We can introduce again the DCT matrix in the last term to get back to cepstral vectors:

$$\begin{aligned} (\mu^c - X^c)^T D^{-T} R D^T \Sigma^{c-1} D R D^{-1}(\mu^c - X^c) = \\ (\mu^c - X^c)^T (DRD^{-1})^T \Sigma^{c-1} \cdot \\ DRD^{-1}(\mu^c - X^c) \quad (4) \end{aligned}$$

Therefore, the only modification to the acoustic models consists to modify the covariance matrix by R as specified in equation 4. The drawback of this method is that original diagonal covariance matrices are now full-covariance matrices.

There are different solutions to estimate R on the training corpus. We have chosen to compute it by normalizing the log-spectral energy of the mean vectors of the Gaussians, so that $R_{imax} = 1$ and $R_{imin} = 1 - \epsilon$, where $imax$ and $imin$ are defined as $imax = \arg\max_i(\mu_i)$ and $imin = \arg\min_i(\mu_i)$. This leads to the following definition:

$$R_i = \frac{\mu_i \cdot \epsilon}{\mu_{imax} - \mu_{imin}} + 1 - \frac{\mu_{imax} \cdot \epsilon}{\mu_{imax} - \mu_{imin}} \quad (5)$$

R can be compared to the "soft masks" of missing data recognition systems [9].

3. Experimental validation

3.1. Experimental setup

29 phones are modeled by 3-emitting-states left-to-right HMMs. Each state contains a mixture of 64 Gaussians. Diagonal covariance matrices are trained, but full-covariance matrices are used during testing after they have been modified by R as described in equation 4. The incoming signal is sampled at 16 kHz and overlapping Hamming windows of 32 ms length are computed for a final frame rate of 10 ms. Each such window is encoded into a 13-coefficients MFCC-CMS vector plus 13 Δ and 13 $\Delta\Delta$ coefficients. The final vector has 39 coefficients.

Tests are realized on the BREF80 database [10], which is the French equivalent of the WallStreet Journal corpus. However, to avoid side-effects due to language models and beaming techniques, we have assessed the proposed method in phonetic recognition only with a null loop-grammar.

In all the following experiments, the noise is added to the clean speech data in the time domain before recognition.

3.2. Experimental results

3.2.1. Band-limited noise

We first test the system on two kinds of artificial band-limited noise: A first noise that affects the [500 Hz ; 600 Hz] frequency range, and another noise that affects the [1500 Hz ; 1800 Hz] frequencies.

Figure 1 and figure 2 respectively plots the phonetic recognition accuracy for the low- and high- frequency noise at an SNR of 10 dB. The x-axis represents the different values of ϵ, while the y-axis represents the phone accuracy. The bold horizontal line represents the baseline accuracy. The thin curve that converges towards this horizontal line on the left represents the corresponding masked models. Note that $\epsilon = 0$ is equivalent to the baseline system.

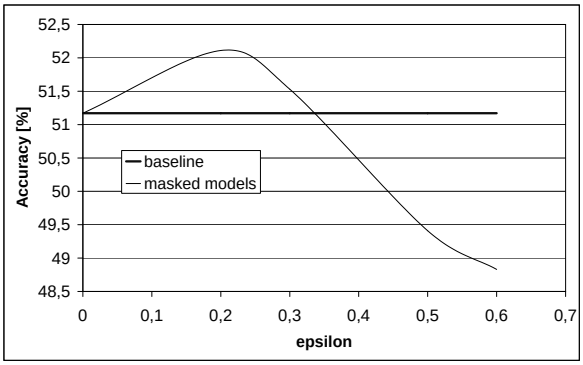

Figure 1: Recognition accuracy of masked models in [500 Hz ; 600 Hz] noise at 10 dB.

We can observe that the proposed masking scheme outperforms the baseline accuracy for a range of ϵ that depends on the type of noise:

- $[0 - 0.3]$ for the low-frequency noise;
- $[0 - 0.6]$ for the high-frequency noise.

However, we can note that for both noises, the optimal ϵ is close to 0.2.

One possible reason for the fact that high-frequency noise is better compensated by the proposed method than low-frequency noise is that frequencies around 500 Hz are more important for speech than frequencies around 1.6 kHz. Therefore, the values of R for speech-like frequencies are close to one and are weakly compensated.

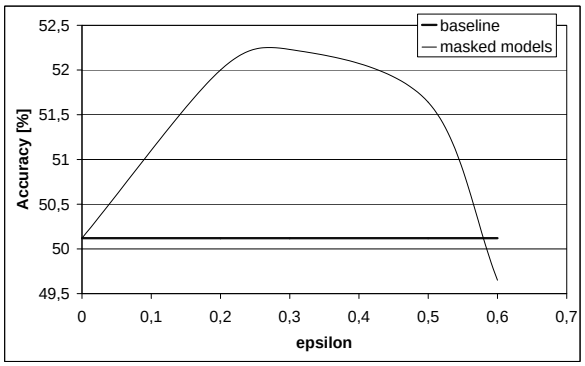

Figure 2: Recognition accuracy of masked models in [1500 Hz ; 1800 Hz] noise at 10 dB.

3.2.2. TV noise

In the following experiments, we use two different non-stationary real-life noises. The first one is a background TV noise. This noise (mainly recorded during commercials) is composed of a main voice along with background music and song: it is clearly non-stationary but also affects a wide range of frequencies. Figure 3 plots the corresponding recognition accuracy.

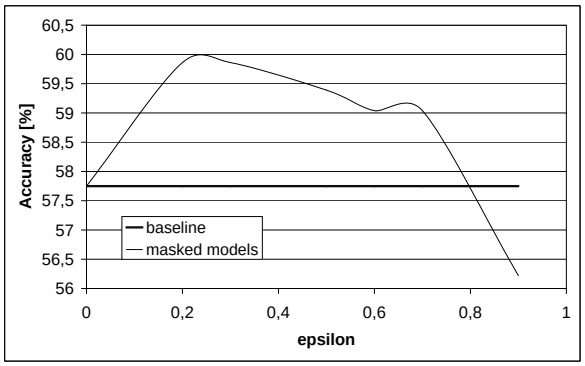

Figure 3: Recognition accuracy of masked models in background TV noise at 15 dB.

We can observe that the improvement of accuracy can be compared to what has been obtained with synthetic noise. Furthermore, the optimal ϵ is still around 0.2.

3.2.3. Musical noise

In the following experiments, Bach's *Chaccone* has been artificially added to the clean test corpus, so as to simulate a background musical noise. It is a solo-violin music. It is therefore a highly non-stationary noise that can hardly be handled by classical adaptation and denoising algorithms. In the following experiment, we assess the proposed method in a range of different SNRs. Figure 4 compares the recognition accuracy of the masked models and of the baseline system at 25 dB, 20 dB, 15 dB, 10 dB and 5 dB.

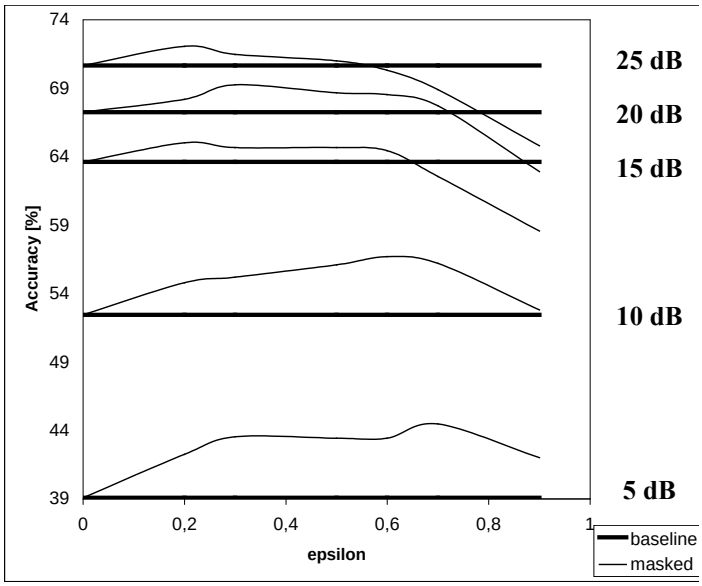

Figure 4: Recognition accuracy of masked models in musical noise at different SNRs

We can conclude from these experiments that the optimal value of ϵ is dependent on the global SNR (and probably on the local SNR as well). Therefore, the algorithm can probably be improved by estimating the global SNR on the test sentences and choosing different ϵ depending on this SNR.

3.2.4. Clean conditions

Modifying the variances of the models is not a good idea in clean conditions, where the models are already optimal. We thus expected some decrease in performances. Surprisingly, the recognition accuracy actually increases, as can be shown in figure 5. We explain this by the fact that the variances of the acoustic models are probably underestimated and that the masking scheme proposed here compensates for this.

4. Discussion

Comparison with missing data recognition masks:

In classical missing data recognition systems, the masks are computed from the estimated signal-to-noise ratio [7] and are the same for all the models. The proposed method rather computes model-specific masks that are independent of the signal. We have derived a "robustness" matrix R that represents the intrinsic robustness of each log-spectral coefficient. This matrix can be com-

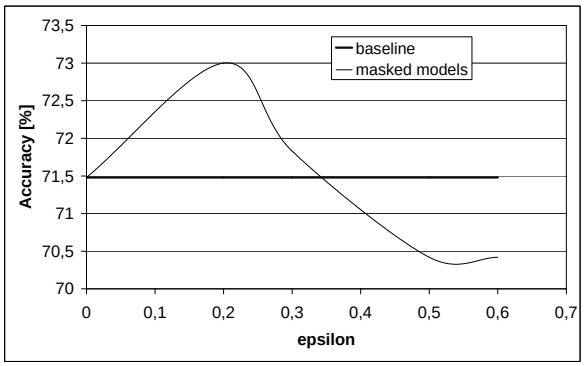

Figure 5: Recognition accuracy of masked models in clean conditions.

pared to the "soft masks" of missing data recognizers, with the possibility to tune the strength of masking via ϵ.

The proposed masking scheme can also be interpreted as a model variance adaptation algorithm: The initial variances trained on the clean corpus are optimal in the Bayes sense when used in clean conditions, but they are not optimal any more in noisy conditions. Our algorithm modifies the variance in the log-spectrum to decrease the importance of the least robust coefficients.

Difference in noise types:

The proposed method is well suited for non-stationary noises, but it does not take into account stationary noises, which affect mainly the means of the Gaussians. Therefore, it should be combined with another noise compensation algorithm that is dedicated to stationary noise, such as spectral subtraction or Parallel Model Combination. On the other hand, it is extremely difficult to accurately identify highly non-stationary noise fragments such as music, using only statistical methods. Thus, we believe that the only way to address the problem is to combine as many sources of information as possible, including the one proposed in this paper.

5. Conclusion

We have proposed an original soft masking scheme in the framework of missing data recognition. The masks are not estimated from the test signal but rather from the acoustic models. The basic principle consists to a priori mask the least robust coefficients of the models. We have validated the proposed method in a variety of noise types at different SNRs. The strength of the method lies in its ability to handle highly non-stationary noise, such as music. However, its performances in quasi-stationary noise are far from the ones obtained with other adaptation methods. We have thus discussed different options to integrate the proposed method into existing robust systems: for example by merging the proposed masks with dynamic masks based on the test signal, or by combining the proposed method with another noise-robust algorithm dedicated to stationary noise. The next objectives consist to address the following issues: (i) Computation cost: full-covariance matrices are much more computationally expensive than diagonal variances; and (ii) Dependency on the SNR: we have observed that the optimal masks are dependent on the global SNR. This might be solved by dynamically estimating the ϵ factor based on the global SNR. Finally, we have designed the proposed approach in order to combine it with other dynamic mask estimation algorithm and we are convinced both approaches will benefit one from the other.

6. Acknowledgements

This work has been partly founded by the IST OZONE project. The authors would like to thank the OZONE project consortium for their support.

7. References

[1] S.T. Roweis, "Factorial models and refiltering for speech separation and denoising," in *EUROSPEECH*, Geneva, 2003, pp. 1009–1012.

[2] H. Bourlard and S. Dupont, "Subband-based speech recognition," in *ICASSP*, 1997, pp. 1251–1254.

[3] C. Cerisara and D. Fohr, "Multi-Band automatic speech recognition," *Computer Speech and Language*, vol. 15, no. 2, pp. 151–174, Apr. 2001.

[4] A. Morris, M. Cooke, and P. Green, "Some solutions to the missing feature problem in data classification, with applications to noise robust asr," in *ICASSP*, Seattle, USA, 1998, pp. 737–740.

[5] B. R. Ramakrishnan, *Reconstruction of incomplete spectrograms for robust speech recognition*, Ph.D. thesis, Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, Pennsylvania, April 2000.

[6] L. Deng, J. Droppo, and A. Acero, "Recursive estimation of nonstationary noise using iterative stochastic approximation for robust speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 6, November 2003.

[7] M. L. Seltzer, "Automatic detection of corrupted speech features for robust speech recognition," M.S. thesis, Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, Pennsylvania, May 2000.

[8] J. Barker, M. Cooke, and D. Ellis, "Decoding speech in the presence of other sound sources," in *ICSLP'00*, Beijing, China, 2000.

[9] J. Barker, L. Josifovski, M. Cooke, and P. Green, "Soft decisions in missing data techniques for robust automatic speech recognition," in *ICSLP'00*, Beijing, China, 2000.

[10] L.F. Lamel, J.L. Gauvain, and M. Eskenazi, "BREF, a large vocabulary spoken corpus for french," in *EUROSPEECH'91*, 1991.