Detection of Reliable Features for Speech Recognition in Noisy Conditions Using a Statistical Criterion

Philippe Renevey[†] *and Andrzej Drygajlo*[‡]

[†] Swiss Center for Electronics and Microtechnology, Neuchâtel, Switzerland
[‡] Swiss Federal Institute of Technology, Lausanne, Switzerland
philippe.renevey@csem.ch, andrzej.drygajlo@epfl.ch

Abstract

This paper addresses the problem of integration of missing data theory in the context of robust speech recognition in additive noise. It shows that techniques based on statistical estimation and thresholding of *a posteriori* signal-to-noise ratio (SNR) can be used for the detection of reliable (not much affected by noise) features as opposed to unreliable or missing (masked by noise) features. In the paper, a statistical detector for reliable features is proposed and tested for several values of deterministic and probabilistic thresholds at very low SNRs (from 20 to -10 dB). The limitations of the detector are also studied and measures for the evaluation of the performance of such a detection are proposed.

1. Introduction

Speech recognition using missing feature approach is based on the assumption that additive noise masks some parts of the time-frequency representation of the speech signal and leaves the other parts not strongly affected. Experiments have shown that the combination of detection of "not very noisy" (*present* or *reliable*) features based on thresholding of *a priori* SNR and recognition processing that uses only these features allows to significantly improve the performance of speech recognizers under noisy conditions [1–3]. Unfortunately, the *a priori* SNR is unavaliable in real operating conditions and detection of the reliable features has to be performed using a sub-optimal criterion. Detection of the reliable features for recognition purposes based on spectral subtraction was first presented by Drygajlo and El-Maliki [4]. Other methods based on a thresholding of *a posteriori* SNR were introduced in [5, 6]. It can be shown that spectral subtraction and SNR based detection are equivalents [7, 8]. The introduction of a soft tresholding (measure of reliability) instead of a hard thresholding allows to improve the performance of speech recognizers [9, 10].

In this paper we present the problem of the detection of reliable features using SNR based detection criteria. We present the theoretical limits of such a detection. Then we study the performances of the detectors based on deterministic and statistical detection criteria.

2. Problem statement

In the approach proposed in this paper the detection of reliable features needs an estimate of the local SNR from the noisy speech features $|Y(\omega,t)|$ and from the noise features $|N(\omega,t)|$. This estimate of the local SNR is then used to divide the features into reliable and unreliable by the use of thresholding.

First we define the local *a priori* SNR as the ratio of the clean signal magnitude to the difference between the magnitudes of the clean and noisy signals:

$$\text{SNR}_{prior}(\omega,t) = \frac{|X(\omega,t)|^2}{\big||Y(\omega,t)| - |X(\omega,t)|\big|^2}. \tag{1}$$

This measure evaluates the distortion introduced by the noise. Unfortunately the clean signal is not available and such a measure cannot be calculated directly.

The addition of two signals in the time domain corresponds to an addition in the spectral domain. The magnitude of the resulting signal depends on the magnitude of the two signals and on the phase difference between the two signals. If $Y(\omega,t)$ is the sum of $X(\omega,t)$ and $N(\omega,t)$, its magnitude can be expressed as:

$$|Y(\omega,t)| = \sqrt{|X(\omega,t)|^2 + |N(\omega,t)|^2 + 2\cdot|X(\omega,t)|\cdot|N(\omega,t)|\cdot\cos(\alpha)} \tag{2}$$

where α is the phase difference between $X(\omega,t)$ and $N(\omega,t)$. The *a priori* SNR defined in Eq. 1 is minimum when , for a given value of $|X(\omega,t)|$, the value of $|Y(\omega,t)|$ is maximum. From Eq. 2 we deduce that $|Y(\omega,t)|$ is maximum when $\alpha = 0$. In this case $|Y(\omega,t)| = |X(\omega,t)| + |N(\omega,t)|$ and Eq. 1 becomes:

$$\text{SNR}_{prior}(\omega,t) = \frac{|X(\omega,t)|^2}{\big||Y(\omega,t)| - |X(\omega,t)|\big|^2} \geq \frac{|X(\omega,t)|^2}{|N(\omega,t)|^2} \geq \left(\frac{|Y(\omega,t)|}{|N(\omega,t)|} - 1\right)^2. \tag{3}$$

The *a posteriori* SNR defined as $\text{SNR}_{post}(\omega,t) = \frac{|Y(\omega,t)|^2}{|N(\omega,t)|^2}$ can be used as a detection criterion because it is directly related to the *a priori* SNR:

$$\text{SNR}_{prior}(\omega,t) \geq \left(\sqrt{\text{SNR}_{post}(\omega,t)} - 1\right)^2. \tag{4}$$

When we have only access to the noisy signal and an estimate of the noise, the selection of the features with an *a priori* SNR higher than a certain value can be obtained by thresholding the *a posteriori* SNR. Therefore, a feature is declared reliable if its *a posteriori* SNR is higher than a threshold value τ. The criterion for the detection of the reliable features is:

$$\text{SNR}_{post}(\omega,t) = \frac{|Y(\omega,t)|^2}{|N(\omega,t)|^2} > \tau. \tag{5}$$

According to Eq. 4 the criterion for the detection of reliable features using Eq. 5 can also be expressed as a function of the *a priori* SNR:

$$\text{SNR}_{prior}(\omega,t) > (\sqrt{\tau} - 1)^2 \tag{6}$$

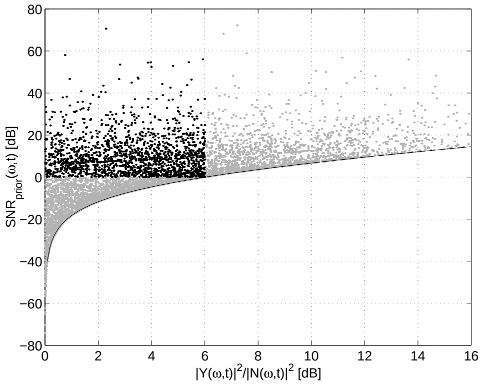

Figure 1: $\text{SNR}_{prior}(\omega,t)$ in function of $\frac{|Y(\omega,t)|^2}{|N(\omega,t)|^2}$ for the addition of two spectra with an unknown phase difference. The black curve corresponds to the lowest possible value described in Eq. 3. The dots are the features. The black dots represent the features which will not be detected as reliable if we fix a threshold $\tau = 6$dB which corresponds to $\text{SNR}_{prior}(\omega,t) \geq 0$ dB. The features are obtained using ten digit utterances with factory noise added at a global SNR of 10 dB.

Fig. 1 presents the *a priori* SNR as a function of the *a posteriori* SNR. The features have been generated artificially by adding artificial signal and noise with a random phase difference. The black curve is the limit of detection defined by Eq. 3. To show the limitations of detection based on the *a posteriori* SNR, we present in this figure an example of detection. We choose to detect as reliable the features with an *a posteriori* SNR higher than 6 dB, which corresponds from Eq. 4 to an *a priori* SNR greater than 0 dB. The black points of Fig. 1 represent the features which have not been detected even if their *a priori* SNRs are higher than 0 dB. This represents the theoretical limit of the SNR based approach for the detection of the reliable features. Even if we estimate the noise spectral magnitude perfectly, some of the reliable features cannot be detected. This is due to the phase difference between the clean speech spectrum and the noise spectrum.

3. Deterministic detection criteria

The first SNR based approach for the detection of reliable features uses a deterministic criterion. The value of the noise magnitude cannot be extracted from the noisy signal. Therefore we use the mean of the noise magnitude calculated during the non-speech segments as the estimate of the noise magnitude. We present a method based on a thresholding of the *a posteriori* SNR and show that other methods based on deterministic criteria, presented in the literature, are similar to the proposed method.

To calculate the local *a posteriori* SNR, the value of the noise signal magnitude $|N(\omega,t)|$ is needed. In a single channel application this value is unavailable and has to be estimated. If we consider that the noise features follow a normal distribution, the mean value $\mu_N(\omega)$ represents the best estimate of the noise in the sense of least square error criterion. The *a posteriori* SNR can therefore be estimated and the feature $|Y(\omega,t)|$ is reliable if:

$$\widehat{\text{SNR}}_{post}(\omega,t) = \frac{|Y(\omega,t)|^2}{|\mu_N(\omega)|^2} > \tau \tag{7}$$

where $\widehat{\text{SNR}}_{post}(\omega,t)$ is the estimate of the *a posteriori* SNR.

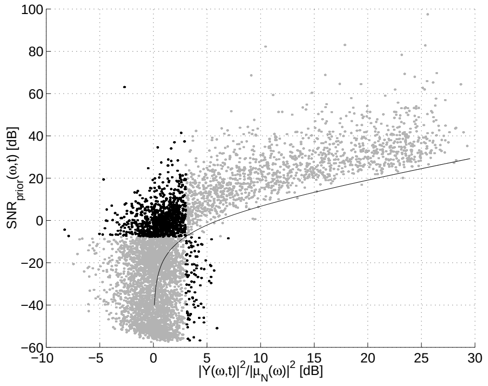

Figure 2: $\text{SNR}_{prior}(\omega,t)$ in function of $\frac{|Y(\omega,t)|^2}{|\mu_N(\omega)|^2}$ for real features. The black curve corresponds to the lowest possible value described in Eq. 3. The dots are the features. The black dots represent the mis-classification errors for a decision threshold τ=3 dB. The features are obtained using ten digit uterances with factory noise added at a global SNR of 10 dB.

Fig. 2 plots the *a priori* SNR as a function of the estimated *a posteriori* SNR. The *a priori* SNR is calculated from the clean and the noisy signal. The *a posteriori* SNR is estimated using the mean of the magnitude of the noise as in Eq. 7. These local SNR values have been obtained using ten sequences of digits extracted from the TIDigit database. Factory noise was added to the features at a global SNR of 10 dB. Misclassified features are plotted with black dots instead of gray dots. The upper-left region represents the features whose *a priori* SNR is higher than the threshold, but which are not detected as reliable. This misclassification is due to the problem of the unknown phase difference between the clean signal and the noise which does not permit correct estimation of the *a priori* SNR. The lower-right misclassification region is due to the fact that the noise magnitude is not equal to its mean but is distributed around this mean. It can therefore be greater than the mean value resulting in a misclassification. The observation of Fig. 2 shows the limit of detection of reliable features based on thresholding the *a posteriori* SNR: as we increase or decrease the value of the decision threshold, one type of misclassification error increases when the other decreases and *vice versa*.

4. Statistical detection criterion

The methods for the detection of reliable features presented above use the average noise magnitude spectrum as the estimation of the magnitude spectrum of the noise. If we consider that the noise magnitude follows a normal distribution in each frequency band, we can represent its distribution in each band by the corresponding mean and variance. The supplementary information of the variance can be introduced into another kind of detector, a statistical detector. This detector calculates a probability for a feature to be reliable rather than to make a hard decision (reliable/unreliable).

In order to estimate the distribution of the noise magnitude in each frequency band, we consider that the noise follows a normal distribution in each sub-band. This distribution of the noise magnitude in each frequency band is therefore expressed as:

$$p\left(|N(\omega)|\Big|\mu_N(\omega),\sigma_N^2(\omega)\right) = \frac{1}{\sqrt{2\pi}|\sigma_N(\omega)|}\exp\left(\frac{(N(\omega)-\mu_N(\omega))^2}{2\sigma_N^2(\omega)}\right). \tag{8}$$

where $\mu_N(\omega)$ and $\sigma_N^2(\omega)$ are, respectively, the mean and variance of the noise in the band ω.

The criterion for the detection of the reliable features presented in Eq. 5 can be rewritten to express the condition of the detection as a function of the noise. In this case, $|Y(\omega,t)|$ is reliable if:

$$|N(\omega)| < \frac{|Y(\omega,t)|}{\sqrt{\tau}}. \tag{9}$$

This threshold can be combined with the model representing the noise magnitude distribution. Therefore we can express the probability that the *a posteriori* SNR is higher then the value τ:

$$\begin{aligned} &P\left(\mathrm{SNR}_{post}(\omega,t) > \tau\Big||Y(\omega,t)|,\mu_N(\omega),\sigma_N^2(\omega)\right) = \\ &P\left(|N(\omega,t)| < \frac{|Y(\omega,t)|}{\sqrt{\tau}}\Big|\mu_N(\omega),\sigma_N^2(\omega)\right) = \\ &\int_{-\infty}^{\frac{|Y(\omega,t)|}{\tau}} \frac{1}{\sqrt{2\pi}|\sigma_N(\omega)|}\exp\left(\frac{(n-\mu_N(\omega))^2}{2\sigma_N^2(\omega)}\right)dn \end{aligned} \tag{10}$$

The hard decision presented previously in Eq. 7 is replaced by a soft decision threshold. In the case of hard decision threshold, a feature can be either reliable or unreliable. The soft decision threshold of Eq. 10 gives a probability between zero and one to be reliable.

To divide explicitly the features into reliable and unreliable, we introduce a threshold value $\theta \in [0,1]$. A feature is declared reliable if:

$$P\left(\mathrm{SNR}_{post}(\omega,t) > \tau\Big||Y(\omega,t)|,\mu_N(\omega),\sigma_N^2(\omega)\right) > \theta \tag{11}$$

where θ represents the probability of having $\mathrm{SNR}_{post}(\omega,t) > \tau$.

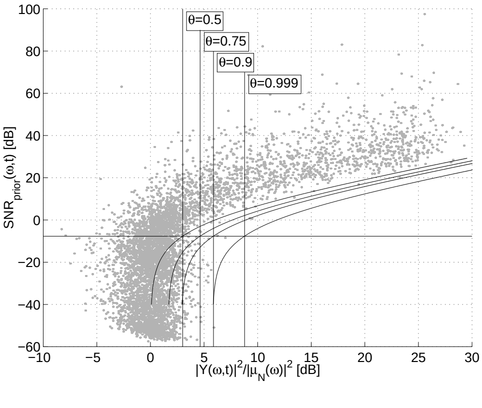

Figure 3: $\mathrm{SNR}_{prior}(\omega,t)$ in function of $\frac{|Y(\omega,t)|^2}{|\mu_N(\omega)|^2}$ for real features. The black curves correspond to the lowest possible value described in Eq. 3 for several values of θ, $\theta = 0.5, 0.75, 0.9, 0.999$. The features are obtained using ten digit uterances with factory noise added at a global SNR of 10 dB.

Fig. 3 plots the *a priori* SNR as a function of the *a posteriori* SNR with the detection thresholds corresponding to $\theta = 0.5, 0.75, 0.9, 0.999$. When $\theta = 0.5$ we have the same threshold as in the deterministic approach presented in Fig. 2. When we increase the value of θ, we decrease the number of unreliable features which are misclassified, but on the other side we increase the number of reliable features which are misclassified. The value of θ allows us to choose between the relative importance of these two errors. If we want to avoid to have unreliable features detected as reliable, we have to choose a higher value for θ (0.99 for example), otherwise, if we want to increase the number of reliable features correctly detected, we will choose a smaller value of θ.

5. Evaluation of the detection

The different detectors presented in the previous section try to divide the features in the spectro-temporal representation of the speech signal in two classes, reliable and unreliable ones, according to a threshold τ. In order to compare the detectors, two measures are defined. These measures compare the proposed detector with an "ideal" detector which uses the clean signal and the noisy signal to compute the *a priori* SNR (Eq. 1) and to divide the features into reliable and unreliable.

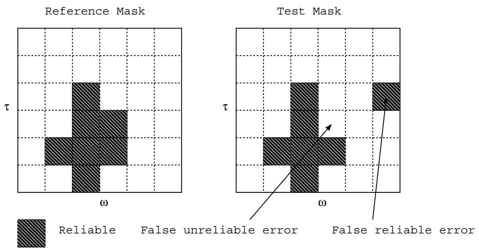

Figure 4: Two possible types of error in reliable/unreliable features detection

For this purpose we define two measure of performance:

- *False unreliable feature detection error (FUFDE):* This error represents the ratio of the number of features that are detected as unreliable when they are reliable to the total number of unreliable features.
- *False reliable feature detection error (FRFDE):* This error represents the ration of the number of features that are detected as reliable when the are reliable to the total number of reliable features.

False unreliable feature detection errors reduce the set of reliable features, but false reliable feature detection errors introduce features masked by noise in the set of reliable features.

6. Performances

The statistical detector for unreliable features presented in Eq. 11 has been tested for several values of θ and τ. The deterministic detectors (Eq. 7) are special cases of the statistical detector, so their performances can be derived from the presented results, using $\theta = 0.5$.

The TIDigit database was used to test the detectors. Sixty-four sentences of seven digits each were extracted randomly. Noises from the Noisex database were added with several global SNRs ranging from -10 to 20 dB. The signals were down-sampled to a frequency of 8 kHz and transformed in the time-frequency domain using a seventeen band Mel filter bank. The reference mask of reliable features, computed using the *a priori* SNR calculated with the clean and the noisy signal (Eq. 1), is compared to those obtained using the proposed detector. The results presented in Fig. 5 represent the measure of the false unreliable feature detection error (FUFDE) as a function of the false reliable feature detection error (FRFDE) (ROC curves) for values of $\tau = 6$ dB.

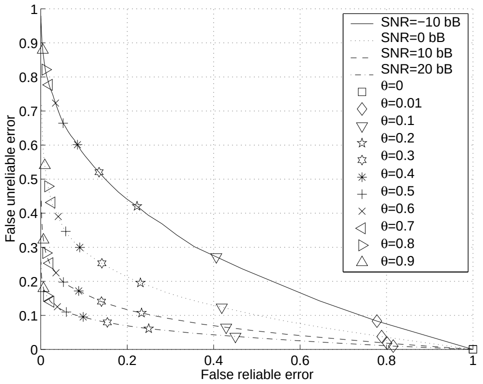

Figure 5: ROC curves for $\tau = 6$ dB and θ value varying from 0 to 1 with white Gaussian noise added with a global SNR varying from -10 to 20 dB by steps of 10 dB.

The value of θ allows a tradeoff between false reliable feature detection errors and false unreliable features detection errors. As the value of θ increases, the value of FUFDE increases and the value of the FRFDE decreases. If the FRFDE is an important feature for the recognition, a high value (near one) of θ has to be chosen.
Several conclusions can be derived from these experiments:

- *As the global SNR increases, the number of misclassification errors decreases.* If the SNR is higher, the difference between speech and noise increases and it becomes easier to detect the reliable features.
- *As the threshold τ increases, the FRFDEs decrease.* As can be observed in Figs. 2 and 3, when the *a posteriori* SNR is small (0 to 5 dB), the FRFDE occur more often than when the *a posteriori* SNR is greater ($>$ 5 dB.).
- *As the threshold τ increases, the number of reliable features correctly detected decreases.* This is explained by the limitation of the chosen detection method presented in Fig. 1.

Recognition results obtained using soft thresholding approach have been presented in [7, 9]. It was observed that the soft thresholding approach allows to improve the performance in several noise conditions (white Gausssian noise, babble noise, Lynx helicopter noise, factory noise, etc.) both for digits and small vocabulary recognition tasks.

7. Conclusion

In this paper we have presented a study of the problem of the detection of the reliable features of a noisy speech signal. We have shown that detection criterion based on *a posteriori* SNR has a theoritical limitation explained by the unknown phase difference between the noise and the speech signal. In real operating conditions, the unknown variation of the noise magnitude also introduces detection errors. A statistical detector has been proposed that allows the tradeoff between two kinds of errors: false reliable- and false unreliable feature detection errors. These errors represent the main limitation of missing feature based approaches developed for the robust speech recognition domain.

8. References

[1] Lippmann, R. P. and Carlson, B. A., "Using missing feature theory to actively select features for robust speech recognition with interruptions, filtering and noise", *in Eurospeech*, vol. 1, pp. 37–40, Rhodes, Greece, Sep. 1997.

[2] Cooke, M., Morris, A., and Green, P., "Missing data techniques for robust speech recognition", *in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 863–866, 1997.

[3] Green, P., Barker, J., Cooke, M., and Josifovski, L., "Handling missing and unreliable information in speech recognition", *in AISTATS*, Florida, USA, 2001.

[4] Drygajlo, A. and El-Maliki, M., "Speaker verification in noisy environments with combined spectral subtraction and missing feature theory", *in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 121–124, 1998.

[5] Vizinho, A., Green, P., Cooke, M., and Josifovski, L., "Missing data theory, spectral subtraction and signal-to-noise estimation for robust asr: an integrated study", *in Eurospeech*, vol. 5, pp. 2407–2410, 1999.

[6] Renevey, P. and Drygajlo, A., "Missing feature theory and probabilistic estimation of clean speech components for robust speech recognition", *in Eurospeech*, vol. 6, pp. 2627–2630, 1999.

[7] Renevey, P., *Speech recognition in noisy conditions using missing feature approach*, PhD thesis, Ecole Polytechnique Fédérale de Lausanne, 2000.

[8] Renevey, P. and Drygajlo, A., "Estimation of unreliable data for robust speech recognition", *in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1731–1734, Istanbul, Turkey, 2000.

[9] Renevey, P. and Drygajlo, A., "Introduction of a reliability measure in missing data approach for robust speech recognition", *in EUSIPCO*, pp. 473–476, Tampere, Finland, 2000.

[10] Barker, J., Josifovski, L., Cooke, M., and Green, P., "Soft decisions in missing data techniques for robust automatic speech recognition", *in Int. Conf. on Spoken Language Processing (ICSLP)*, 2000.