Mask Estimation in Non-stationary Noise Environments for Missing Feature Based Robust Speech Recognition

Shirin Badiezadegan and Richard C. Rose

Department of Electrical and Computer Engineering, McGill University, Canada
shirin.badiezadegan@mail.mcgill.ca, rose@ece.mcgill.ca

Abstract
In missing feature based automatic speech recognition (ASR), the role of the spectro-temporal mask in providing an accurate description of the relationship between target speech and environmental noise is critical for minimizing the degradation in ASR word accuracy (WAC) as the signal-to-noise ratio (SNR) decreases. This paper demonstrates the importance of accurate characterization of instantaneous acoustic background for mask estimation in data imputation approaches to missing feature based ASR, especially in the presence of non-stationary background noise. Mask estimation relies on a hypothesis test designed to detect the presence of speech in time-frequency spectral bins under rapidly varying noise conditions. Masked mel-frequency filter bank energies are reconstructed using a minimum mean squared error (MMSE) based data imputation procedure. The impact of this mask estimation approach is evaluated in the context of MMSE based data imputation under multiple background conditions over a range of SNRs using the Aurora 2 speech corpus.

Index Terms: Automatic speech recognition, Missing feature techniques, soft mask estimation, spectrogram reconstruction.

1. Introduction
Acoustic robustness in automatic speech recognition (ASR) implies the ability to maintain a high level of recognition accuracy in the presence of multiple sources of variability, especially when the recognizer has been trained under noise-free conditions [1], [2]. There have been a large number of feature compensation and model adaptation approaches that have demonstrated good performance in environments where the main source of acoustical variability is stationary noise with moderate SNR [3]. However, most scenarios of general interest, like, for example, ASR services designed for mobile users and voice control of devices from far field microphones, can be characterized by rapidly varying non-stationary noise conditions.

This paper addresses the issue of non-stationary background noise within a missing feature framework. Missing feature techniques first segment sound sources in the time-frequency domain into “reliable” regions which are dominated by the target speech and “missing” or “unreliable” regions which are dominated by background noise [4]. This segregation is represented by a spectro-temporal mask which assigns a probability of speech presence to individual time-frequency spectral components. In the data imputation approach to missing feature based ASR, an estimate is obtained for the spectral components labeled as unreliable. ASR features are computed from the reconstructed spectral representation, and these features are used in an ASR system which has been trained in a noise-free environment. In this work, a minimum mean squared-error (MMSE) based data imputation approach is used to re-estimate the unreliable spectrogram components [2], [5]. This MMSE approach is described in Section 2.

The purpose of this paper is to develop robust methods for estimating noise masks in non-stationary acoustic environments for data imputation based missing feature approaches. These methods are based on techniques, originally developed for speech enhancement applications, for estimating the characteristics of speech and background information in noisy speech utterances [6]. In these applications, the importance of obtaining accurate local estimates of rapidly varying noise characteristics and the probability of speech presence in time-frequency spectral components is well understood [6]. For example, the accuracy of this information is known to be critical for achieving good performance in minimum mean squared error - log spectral amplitude (MMSE-LSA) based speech enhancement. Section 3 describes a mask estimation procedure which is based on a hypothesis test designed to detect the presence of speech in spectral components under rapidly varying noise conditions.

There is a large body of literature dealing with mask estimation in a variety of sometimes extremely challenging acoustic environments [7]. However, many of these techniques make very specific assumptions about the characteristics of the background conditions. A common assumption is that the acoustic background is a stationary process and that the mask can be implemented using a parametric model of background. The parameters of this model are trained prior to recognition from signals that are thought to be representative of the target background environment [5],[7]. For these techniques to be effective in general background environments, it is necessary to allow for variation in both the characteristics of the noise and the noise level in the target environment. Adapting to these new noise conditions can be extremely difficult when parametric models of acoustic background are used. The proposed mask estimation method, on the other hand, is specifically designed to identify the speech and noise dominant spectrogram components in the presence of rapidly varying non-stationary noise.

An experimental study is presented in Section 4 to evaluate the impact of this mask estimation approach in the context of MMSE based data imputation under multiple background conditions over a range of SNRs using the Aurora 2 speech corpus. It will be shown in many cases to out-perform mask estimation approaches that rely on prior knowledge of noise type and SNR.

2. MMSE Approach to Data Imputation
This section describes a brief summary of a MMSE approach to data imputation based missing feature estimation. The speech corruption model will be described along with a description of how this corruption model impacts the component Gaussians
in a Gaussian mixture model of the underlying clean speech. This model is then used to obtain the MMSE estimate of clean speech spectral components from noise corrupted observations. Missing feature techniques generally assume that the spectral energy time-frequency bins have been partitioned into reliable and unreliable portions by a spectro-temporal mask. Data imputation techniques estimate the values of unreliable components based on their statistical relationship with the reliable components resulting in a set of reconstructed spectral energies. This procedure is an alternative to marginalization approaches to missing feature based ASR where information provided by the mask is used to modify the probability densities in the ASR system [3]. A comparison of the relative advantages of these two approaches is beyond the scope of this paper, and only data imputation will be considered in the following discussion.

2.1. Model for Speech Corruption

Data imputation methods are based on the assumption that target speech and background noise are additive and uncorrelated in the linear spectral domain. The effect of noise in the log spectral domain or the mel-warped log spectral domain is approximated as occlusion of the clean speech features, indicating that each spectral component is either dominated by speech or dominated by background noise [5,8]. If we denote noise corrupted speech, clean speech, and the background noise spectra in the log mel-spectrum domain by $y$, $x$, and $w$, respectively, the noise corrupted speech can be expressed as:

$$y_d = \max(x_d, w_d), \quad d = 1, \ldots, D,$$

in which $d$ is the index of the $D$ component mel-scale filter-bank vector. We also denote the soft mask probability corresponding to $y_d$ by $\theta_d$. Considering the above occlusion model:

$$\theta_d = P(x_d \geq w_d),$$

(2)

While there is no parametric form assumed for $\theta_d$, it is often assumed that the uncorrupted speech can be described by a mixture of Gaussians,

$$P(x) = \sum_k c_k N(x; \mu_k, \Sigma_k),$$

(3)

where $c_k$ are mixture weights, and $N(x; \mu_k, \Sigma_k)$ is a normal density with mean $\mu_k$ and covariance matrix $\Sigma_k$. It is a common practice to assume that the covariance matrices are diagonal in order to reduce the computational complexity, so that $\Sigma_k = \text{diag}(\sigma_k^2_1, \ldots, \sigma_k^2_D)$.

The noise corruption model given in Equation 1 has the effect of restricting the domain of $x_d$, so that $x_d \leq y_d$. It is well known that if $x_d$ is normally distributed, then $x_d$ conditioned on $L_d \leq x_d \leq U_d$ has a truncated normal distribution, $N^{[L_d, U_d]}$, with mean $\mu_k^{[L_d, U_d]}$ given by:

$$\mu_k^{[L_d, U_d]} = \mu_k - \sigma_k \frac{N(U_d; \mu_k, \sigma_k^2) - N(L_d; \mu_k, \sigma_k^2)}{C(U_d; \mu_k, \sigma_k^2) - C(L_d; \mu_k, \sigma_k^2)},$$

(4)

where $L_d$ and $U_d$ are the lower and upper bounds on the value of $x_d$ and $C$ is the cumulative Gaussian distribution. It will be shown in Section 2.2 that the effect of the data occlusion model in Equation 4 is to introduce the truncated Gaussian mean values as given in Equation 4 when replacing the speech spectral components with their MMSE estimates.

2.2. MMSE-based Spectrogram Reconstruction

A brief discussion is provided here of the MMSE estimation of clean speech following the developments given in [5], [8]. The MMSE-based estimate of clean speech component, $x_d$, given the noisy speech, $y$, and mask $\theta = [\theta_1, \ldots, \theta_D]$ is defined as:

$$\hat{x}_d = E[x_d|y, \theta].$$

(5)

If $x$ is modeled as Gaussian mixture in Equation 3, then it can be shown that:

$$E[x_d|y, \theta] = \sum_k p(k|y, \theta)E[x_d|y, \theta, k],$$

(6)

where $p(k|y, \theta)$ is the posterior probability of mixture component $k$ and

$$E[x_d|y, \theta, k] = \int_{L_d}^{U_d} x_d p(x_d|y, \theta, k) dx_d.$$  

(7)

In Equation 7, $y_d$ is the upper bound on $x_d$ according to the occlusion model and $L_d$ is an empirically determined lower bound.

According to the definition of spectrographic soft mask and the occlusion model stated in Equations 1 and 2, the component $y_d$ is reliable with the probability $\theta_d$, i.e. $y_d = x_d$, and $y_d$ is not reliable with the probability $1 - \theta_d$, i.e. $y_d = w_d$, and $x_d < y_d$. Thus, it can be stated that:

$$p(x_d|y, \theta, k) = \theta_d \delta_{y_d}(x_d) + (1 - \theta_d) N^{[L_d, U_d]}(x_d; \mu_k, \sigma_k^2),$$

(8)

where $\delta_{y_d}$ is a Dirac-delta function centered at $y_d$. The MMSE estimate of $x_d$ can be obtained from Equations 6, 7, and 8 as:

$$\hat{x}_d = \theta_d y_d + (1 - \theta_d) \sum_k p(k|y, \theta)\mu_k^{[L_d, U_d]}.$$  

(9)

The first term in Equation 9 models the scenario where no occlusion has occurred and the second term corresponds to the scenario where the observed value is not reliable and the underlying clean speech feature should be estimated as a weighted sum of the truncated Gaussians’ expected values. The parametrization of the model and its application to robust ASR will be described in detail in Section 4.

3. Spectrographic Soft Mask Estimation

This section describes a procedure for estimating spectro-temporal masks using a hypothesis test for detecting the presence of speech in time-frequency spectral bins. This will be referred to as the speech presence probability (SPP) mask. For the purposes of comparison, a model based soft mask estimation algorithm which was proposed in [9] and used in [5] is briefly summarized. This will be referred to as the Gaussian mixture model (GMM) mask. It is important to note that the SPP mask is designed to continually update estimates of background level, SNR, and speech presence probability while relying on no prior information concerning the background conditions. The GMM mask as it is implemented here is trained from noise signals of the same type and level that exist in the recognition utterances.

3.1. A Hypothesis Test for Detecting Speech Presence

A procedure similar to the one proposed in [6] is used here for estimating the probability of speech presence in each time frame-frequency bin. It is assumed that the DFT coefficients
of speech and noise are statistically independent Gaussian variables in each frame. This assumption is valid even for a non-stationary noise since the analysis window of the feature extraction procedure is about 20-30 ms.

Assuming that \( y(n), x(n) \), and \( w(n) \) are the noisy speech, the clean speech, and the background noise, the noise corruption can be stated as:

\[
y(n) = x(n) + w(n).
\]

The \( k \)th DFT coefficients for \( x(n) \) and \( y(n) \) are defined as:

\[
X_k = A_k \exp(j\phi_k), \quad Y_k = R_k \exp(j\theta_k),
\]

and are assumed to be statistically independent Gaussian random variables. In [6] the term a-posteriori SNR, \( \gamma_k \), has been defined for the \( k \)th frequency bin as:

\[
\gamma_k = \frac{R_k^2}{\lambda_w(k)},
\]

where \( \lambda_w(k) = E[|W_k|^2] \), and \( |W_k|^2 \) is the noise spectral energy. It is important to note that the value of \( \lambda_w(k) \) is continually updated by using minimum statistics of the smoothed DFT to track the noise variations. To determine whether speech is present in a given frequency bin of a given time frame, Malah et al. pose the following hypothesis testing problem in which the null hypothesis, \( H_0 \), is the presence of speech [6]:

\[
H_0 : \gamma_k > \gamma_{TH}
\]

\[
H_1 : \gamma_k < \gamma_{TH}
\]

The above hypothesis test is repeated for each frequency bin within each time frame. If \( H_0 \) is rejected for the \( k \)th bin of the \( l \)th frame, an index function \( I_k(l) \) is set to 1, and \( I_k(l) = 0 \) if \( H_0 \) is accepted. To estimate the a-priori probability of speech absence in the \( k \)th bin, \( q_k \), the following estimator is proposed in [6]:

\[
\tilde{q}_k(l) = a_0 \tilde{q}_k(l-1) + (1 - a_0) I_k(l).
\]

The values of the parameters \( a_0 \) and \( \gamma_{TH} \) need to be set empirically. In our experiments, an improvement in performance of the ASR system was observed with \( a_0 = 0.5 \) and \( \gamma_{TH} = 6.26 \), which ensures a low probability of false alarm.

The speech absence probabilities is Equation 14 are estimated for each DFT coefficient, \( Y_k \). It is necessary to combine these estimates to obtain the mask probability \( \Theta_d \) as obtained in Equation 2 for mel log filter-bank output \( y_d \). The mel-scale filter-banks are linearly spaced in a non-linear mel-frequency scale, where filter \( d \) covers \( N_d \) frequency bins. To estimate the probability of speech presence for filter output \( d \), we need to calculate the probability of the weighted sum of \( N_d \) independent variables, \( \Theta_{d} = \sum_{i=1}^{N_d} a_{i,d} B_{i} \), where each random variable, \( B_i \), is the presence of speech in a given time-frequency bin and \( a_{i,d} \)'s are the coefficients of the triangular weighting functions associated with the mel filter-bank satisfying \( \sum_{i=1}^{N_d} a_{i,d} = 1 \). Each random variable \( B_i \) can be defined as a Bernoulli random variable with success probability \( p \), where \( p = 1 - q_i(l) \). We use the expected value of \( \Theta_{d} \) as an estimate of the probability of speech presence for each mel filter output. Therefore, \( \tilde{\Theta}_{d} \), the mask value for the \( i \)th time frame and \( d \)th mel filter index is given by:

\[
\tilde{\Theta}_d = E[\Theta_d] = \sum_{i=1}^{N_d} a_{i,d} E[B_i] = \sum_{i=1}^{N_d} a_{i,d}(1 - q_i(l)).
\]

### 3.2. GMM-Based Soft Mask Estimation

This method is based on the occlusion model described in Section 2.1. It relies on the assumption that the clean speech and noise mel log features can be modeled by Gaussian mixtures, the parameters of which are assumed to be known prior to mask estimation [5,9]. The soft mask component \( \Theta_{d} \), corresponding to \( y_d \), reflects the probability that \( x_d > w_d \). With GMM models used for clean speech and noise features, this probability can be given by:

\[
\theta_d = \sum_{k,j} \frac{p(k,j)p_y(y_d|k)C_w(y_d|k)}{C_y(y_d|k) + p_x(y_d|k)C_w(y_d|k)}.
\]

where \( p_y \) and \( p_w \) are the Gaussian mixture components of the clean speech and noise GMMs and \( C_y \) and \( C_w \) are the corresponding cumulative Gaussian distributions. In Equation 16, the model parameters are dependent on both the noise type and the noise level. This information is assumed to be known when this particular mask estimation procedure is implemented for the experimental study described in Section 4.

### 4. Experimental Study

The goal of this study is to demonstrate the importance of continually updated estimates of instantaneous speech and background for use in noise masking as provided by the SPP based mask estimation procedure. Both the SPP and GMM based mask estimation procedures will be evaluated in the context of the MMSE based data imputation approach described in Section 3. The study will compare the SPP approach which is implemented with no prior knowledge of background conditions, and the parametric GMM approach which is trained from noise signals of the same type and level as used for ASR.

#### 4.1. Task Domain and Implementation

All approaches were evaluated on the Aurora 2 connected digit task domain. Test set "a" in the Aurora 2 corpus with subway and babble noise were used for evaluation, which are instances of stationary and non-stationary noises, respectively. In the Aurora2 corpus, noisy utterances are formed by combining noise signals with clean speech utterances of connected digit sequences. The conditions represented in the corpus are somewhat limited since the average SNR for each test utterance is relatively constant. However, the availability of multiple noise types at multiple SNRs is valuable for evaluating the ability of the SPP mask estimation approach to generalize to unseen environments. ASR feature analysis was performed by extracting mel log spectral features using a 25 ms Hamming window, updated every 10 ms. A 512-point FFT was applied to evaluate the spectral values, and a mel-scale filter-bank with 23 filters spaced on a mel frequency scale was used to generate the mel log spectral features.

Whole word digit models were trained using the Aurora 2 clean speech training set and the standard model parameterization and system configuration specified for the Aurora 2 evaluation were used in this study [10]. To reduce the computational load and memory requirements in missing feature estimation, the mel-log spectral features were quantized as 8-bit integers since they have limited dynamic range. An ASR WAC of 98.88% was obtained on the clean speech test set for the above system configuration. It was found that M=40 mixture components were sufficient for the GMM speech model in Equation 3 which is far smaller that the M=512 components used in [5].
The performance for all the systems evaluated in the study is reported as ASR word accuracy (WAC) over SNRs ranging from 5 to 20 dB for the subway noise condition in Table 1 and the speech babble noise condition in Table 2. The first row in Tables 1 and 2 displays the baseline ASR performance obtained when no feature compensation is performed. The degradation in WAC as the SNR level is reduced is consistent with results reported elsewhere for this task [10].

4.2. Mask Estimation and ASR Performance

A first experiment was performed to determine the best case performance that can be achieved by mean imputation based spectrum reconstruction assuming a near perfect mask. An ideal mask for each noisy utterance at each SNR level was obtained when combining the separate speech and noise files. Masks were computed by comparing clean speech spectrum levels to noise signal spectral levels for each time-frequency bin when the noise signals were scaled for a given average SNR. The performance for MMSE based spectrum reconstruction using this ideal mask is displayed in the second row of Tables 1 and 2. It is clear from the tables that the WAC obtained at all SNR levels using the ideal mask is remarkably high. In fact, even at SNR levels as low as 5 dB, the WAC approaches that obtained for clean conditions. This is consistent with similar experiments showing that missing feature techniques perform extremely well given near perfect knowledge of how speech and background interact [3].

A second experiment was performed using the GMM based mask estimation described in Section 3.2. A separate set of model parameters were trained for each of the two noise conditions and for all four SNR levels. Hence, the performance reported for the GMM mask in the third row of Tables 1 and 2 represents the condition where there is prior knowledge of both the noise type and the SNR.

The performance of the SPP based mask estimation approach described in Section 3.1 was evaluated under the assumption that there is no prior knowledge of noise type or SNR. The ASR WAC is given in the fourth row of Tables 1 and 2. It is interesting to compare the performance obtained for the SPP and GMM masks. It is clear from the results that the relative comparison between the two masks is dependent on the noise type. For the subway noise type shown in Table 1, the performance of the two are relatively similar, except for the 5 dB SNR condition. For the speech babble noise type shown in Table 2, the WAC obtained for the SPP mask is consistently higher than that obtained for the GMM mask across all SNR conditions.

The most likely explanation for this behavior are the differences in the characteristics of the two noise types. The subway noise type is generally considered to be stationary. In that case, the GMM mask’s prior estimates of the model parameters may be more important than the SPP mask’s ability to track instantaneous background characteristics. The babble noise is generally considered to be non-stationary. The SPP mask’s ability to track the effects of this non-stationary background is shown to provide a significant advantage in this case despite the prior information used by the GMM mask.

5. Conclusion

A speech presence probability (SPP) approach for estimating spectrographic masks in missing feature based robust ASR has been presented. The important attributes of the approach were its ability to continually update estimates of background level, SNR, and speech presence probability while relying on no prior information about the characteristics of the background noise. This approach was evaluated in the context of an MMSE based data imputation procedure. Experimental results obtained for the Aurora 2 task domain demonstrated that, for non-stationary background conditions, the SPP mask outperforms the model based GMM mask even when the GMM mask is estimated with prior knowledge of noise type and SNR.

6. References


| Table 1: ASR WAC for Data Imputation on Aurora2, test set a, subway noise. |
|-----------------|-----|-----|-----|-----|-----|
| Method          | 5 dB| 10 dB| 15 dB| 20 dB|
| Baseline        | 55.36| 79.21| 92.23| 96.19|
| Ideal Mask      | 91.34| 96.19| 97.88| 97.91|
| GMM Mask        | 70.34| 84.19| 91.43| 96.19|
| SPP Mask        | 66.10| 83.21| 92.29| 95.21|

| Table 2: ASR WAC for Data Imputation on Aurora2, test set a, babble noise. |
|-----------------|-----|-----|-----|-----|-----|
| Method          | 5 dB| 10 dB| 15 dB| 20 dB|
| Baseline        | 19.44| 52.24| 81.98| 92.39|
| Ideal Mask      | 95.74| 97.40| 98.31| 98.34|
| GMM Mask        | 58.46| 69.95| 75.64| 92.39|
| SPP Mask        | 70.22| 89.51| 95.31| 97.61|