SPEECH ENHANCEMENT BASED ON LOG SPECTRAL ENVELOPE MODEL AND HARMONICITY-DERIVED SPECTRAL MASK, AND ITS COUPLING WITH FEATURE COMPENSATION

Takuya Yoshioka, Tomohiro Nakatani

NTT Communication Science Laboratories, NTT Corporation
2-4, Hikari-dai, Seika-cho, Soraku-gun, Kyoto 619-0237, Japan

ABSTRACT

The use of a speech spectral envelope model defined in the log spectrum-type domain is a common approach to feature enhancement for noise robust speech recognition. However, from the noise reduction viewpoint, this approach ignores non-peak components of a spectrum and thus suffers from the poor SNR improvement during voiced periods. This paper proposes a speech enhancement method that exploits a log spectral envelope model and a harmonic structure. The key to the method is the use of a harmonic structure to define the prior distribution of a spectral mask, which is used for both accurate noise estimation and attenuation. In addition, we combine log mel-frequency feature enhancement with the above method to take advantage of low dimensionality. The whole proposed method outperforms a state-of-the-art speech enhancement method in four different noise environments.

Index Terms— Speech enhancement, log spectrum, log mel-frequency spectrum, spectral mask, harmonic structure

1. INTRODUCTION

Speech signals are highly structured due to physical and linguistic constraints, and separating speech from acoustic noise by using such speech characteristics is a fundamental goal of speech enhancement. For example, the short time spectrum of a speech signal exhibits a strong harmonic structure during voiced segments; speech energy is sparsely distributed along frequency; and the possible patterns of log spectral envelopes are limited. All of these characteristics can be and have been used for speech enhancement in many different ways.

There are two previously proposed approaches to speech enhancement: one involves prior training of a speech model while the other does not. OMLSA (Optimally Modified Log Spectral Amplitude estimator) combined with the IMCRA (Improved Minima Controlled Recursive Averaging) noise estimator is one of the state-of-the-art techniques that adopts the no-prior-training approach and it considerably improves SNRs without producing musical noise in stationary or slowly varying noise environments [1, 2]. The sparsity of speech is capitalized on in these recent methods, where the sparsity is used to estimate a speech presence probability vector (or sometimes called a spectral mask).

The other approach, which is based on the prior training, is characterized by the use of a spectral envelope model of speech to restore the overall spectral shape effectively. A codebook of an all-pole spectrum [3] and a GMM of a log mel-frequency spectrum [4] are often used for the speech modeling. Although the latter is usually used for feature enhancement, it can be used to yield an enhanced speech signal via mel-warped IDCT. The log mel-frequency feature GMM generally better represents the distribution of a spectral envelope than the all-pole spectrum codebook as is known in the speech recognition field. Our informal experiment showed that the feature GMM yielded better speech enhancement results.

The main downside to the prior training-based approach is residual noise synchronized with voice activity. Specifically, a significant portion of noise tends to remain at the spectral valleys between harmonic peaks during voiced periods. This is because the spectral components at non-peak frequencies have little effect on the overall spectral shape. The method proposed in [5] directly models a log spectrum without making any distinction between the envelope and the fine structure of a spectrum, thereby avoiding this problem. However, this simple modeling policy forces us to use a GMM with lots of mixture components.

With this as the background, in this paper, we propose a new speech enhancement method, called SPEAC (Spectrum and FEature Combination). SPEAC consists of three components: a harmonic structure model, a log spectral envelope model represented directly in the log spectrum domain, and a log mel-frequency feature model. The harmonic structure model and the log spectral envelope model are used in a unified way for log spectrum enhancement (Section 3). The key to the method is its use of a harmonic structure for defining the prior distribution of a spectral mask, in which various F0 candidates are explicitly considered. This enables us to obtain a posterior distribution of a spectral mask while estimating a spectral envelope of clean speech and an F0 at the same time. Moreover, this posterior distribution is also used to update the noise model. On the negative side, the high dimensionality of a log spectrum representation makes the method sensitive to model mismatch. In order to avoid this, SPEAC combines the log spectrum enhancement method described in Section 3 with the log mel-frequency feature enhancement (Section 4) to take advantage of low dimensionality. The feature model is also employed for voiced/unvoiced classification, which is used to determine whether to enable or disable the harmonic structure model. Large-scale experimental results showed that SPEAC outperformed OMLSA-IMCRA especially in highly non-stationary noise environments (Section 5).

2. MODEL-BASED FEATURE ENHANCEMENT

We begin by reviewing a model-based approach to the enhancement of a feature vector such as a log mel-frequency spectrum. The proposed method will be developed based on this approach. For a full description of this approach, we refer the reader to [4].

Here, we assume that speech is represented in the log mel-frequency feature or log spectrum domain, in which the noise mixing process is assumed to be independent for each dimension. Let \( z = \)
\[ x_1, \ldots, x_L^T, y = [y_1, \ldots, y_L^T], \text{and } n = [n_1, \ldots, n_L^T] \] denote short time frames of clean speech, noisy speech, and noise, respectively. \( L \) is the number of mel-frequency channels or frequency bins.

The central task in feature enhancement and speech enhancement is to obtain the posterior distribution of the clean speech \( p_{SV}(x|y) \) given the noisy speech. An estimate \( \hat{x} \) of the clean speech is obtained by taking the mean of \( p_{SV}(x|y) \). To obtain an enhanced speech signal, we construct a time-domain noise reduction filter from \( w = \exp(\hat{x} - y) \) and apply it to a noisy speech signal. The time-domain filter is calculated as \( g_{\text{MDCT}}(w) \) (when the log mel-frequency feature representation is used) or \( g_{\text{IDCT}}(w) \) (when the log spectrum representation is used), where \( g_{\text{MDCT}} \) and \( g_{\text{IDCT}} \) denote mel-warped IDCT and IDCT, respectively.

We model a prior distribution of speech with a GMM with \( K \) components. Thus, we have
\[
p_{x}(x) = \sum_{k=1}^{K} p_{x}(k) \left( \prod_{i=1}^{L} p_{ux}(x_i|k) \right) = \sum_{k=1}^{K} \prod_{i=1}^{L} \left( f_{x}(x_i; \mu_i, \sigma_i) \right),
\]
where \( f_{x}(x_i; \mu_i, \sigma_i) \) is the pdf of a Gaussian distribution with mean \( \mu_i \) and variance \( \sigma_i \). The speech model is trained in advance. Note that the conditional independence of the vector elements is assumed in (1). When a noise model \( p_{n}(n) \), we assume a single Gaussian distribution and denote its mean and variance by \( \mu_n^2 \) and \( \sigma_n^2 \), respectively.

Then, the posterior pdf of clean speech is calculated as
\[
p_{SV}(x|y) = \sum_{k=1}^{K} p_{SV}(k|y) \left( \prod_{i=1}^{L} p_{ux}(x_i|k) \right).
\]

We calculate the posterior probability of a mixture component \( p_{SV}(k|y) \) and the conditional pdf of speech \( p_{SV}(x|y, k) \) according to the Bayes theorem. To do this, we need a noise model \( p_{n}(n) \) and a noisy speech model \( p_{SV}(y|x,k) \), which are defined as described below.

The noise mixing model is derived by using the first-order VTS (Vector Taylor Series) approximation of the well-known noise mixing formula \( y_i = x_i + \log(1 + \exp(n_i - x_i)) \) [1]. That is, we approximate linearize this equation with respect to (w.r.t.) \( x_i \) and \( n_i \) as
\[
y_i \approx a(x_i^0, n_i^0)x_i + b(x_i^0, n_i^0)n_i + c(x_i^0, n_i^0),
\]
where \( [x_i^0, n_i^0]^T \) is the linearization point, \( a \) and \( b \) are the first derivatives w.r.t. \( x_i \) and \( n_i \), respectively, and \( c \) is the zeroth-order term of the approximation. Assuming that the linearization point consists of the speech and noise means and regarding \( n_i \) as a random variable, we obtain the noise mixing model as
\[
p_{SV}(y|x_i, k) = f_{x}(y_i; \mu_{ux}^k, \sigma_{ux}^k),
\]
\[
\mu_{ux}^k = a(x_i^0, n_i^0)x_i + b(x_i^0, n_i^0)n_i + c(x_i^0, n_i^0),
\]
\[
\sigma_{ux}^k = b(x_i^0, n_i^0)^2n_i^2.
\]

Furthermore, the posterior speech model is easily obtained by taking the expectation of \( p_{SV}(y|x_i, k) \) over \( x_i \) as
\[
p_{SV}(y|k) = \int p_{SV}(y|x_i, k)p_{SV}(x_i|k)dx_i = f_{x}(y_i; \mu_{ux}^k, \sigma_{ux}^k),
\]
\[
\mu_{ux}^k = a(x_i^0, n_i^0)x_i + b(x_i^0, n_i^0)n_i + c(x_i^0, n_i^0),
\]
\[
\sigma_{ux}^k = b(x_i^0, n_i^0)^2n_i^2.
\]

By using equations (1), (4), and (7) in (2), the posterior pdf \( p_{SV}(x|y) \) of clean speech is calculated.

3. SPEECH ENHANCEMENT BASED ON LOG SPECTRAL ENVELOPE MODEL AND HARMONICTY-DERIVED SPECTRAL MASK

In this section, we propose a speech enhancement method that capitalizes on a harmonic structure along with a log spectral envelope model. This method constitutes the log spectrum enhancement part of SPEAC.

This method uses a speech model consisting of two components (see Fig. 1 for comparison of the models used in Sections 2–4):

- **log spectral envelope** — which is modeled by a GMM acquired by prior training;
- **spectral mask** — each element of which indicates the presence/absence of voiced sound at a respective frequency bin. A prior distribution of the spectral mask is modeled by a mixture of harmonically constrained multivariate Bernoulli distributions as we will see later. Each harmonically constrained multivariate Bernoulli distribution corresponds to one F0 candidate.

This speech model lets us estimate the clean speech spectral envelope, the posterior distribution of a spectral mask, and the F0 simultaneously. The posterior distribution of the spectral mask is used both to improve the SNR and to update the noise model as described in Section 4.1.

3.1. Models of speech and noise mixing process

Now we define the speech model and the noise mixing model used in the proposed method in reference to Fig. 1 (b).

We denote a log spectral envelope of clean speech and a spectral mask by \( x = [x_1, \ldots, x_L^T] \) and \( z = [z_1, \ldots, z_L^T] \), respectively, where \( L \) is the number of frequency bins, Each \( z_i \) is a binary-valued variable, \( z_i = 1 \) indicates that the log spectral component of speech at the corresponding frequency bin is \( x_i \), By contrast, \( z_i = 0 \) means the absence of speech at that frequency bin. We assume that the prior distribution of the log spectral envelope is given by a GMM as in (1).

The first key to the proposed model lies in the definition of the noise mixing model for the log spectrum representation. The new noise mixing model is described as
\[
p_{SV}(x|z_i, k) = p_{SV}(y|x_i, k)^{z_i}p_n(y)^{1-z_i},
\]

![Fig. 1. Noisy speech generation models for Sections 2–4. Newly added parts are encircled. In (c), GMM component index \( k \) is used as a junction of the log spectrum and feature domains.](image)
where $p_{y_{1:k}}$ is given by (4). This model suggests that the observed spectrum components are mixtures of speech and noise at frequency bins with $z_i = 1$ while the observed spectral components consist of noise at frequency bins with $z_i = 0$.

The second key concerns the prior distribution of the spectral mask $z$. Because $z_i$ is likely to be 1 only around harmonic frequencies, we wish to define it so that harmonic structures for various F0 candidates are taken into account. For this purpose, we propose a mixture of harmonically constrained multivariate Bernoulli distributions defined as

$$p_z(z) = \sum_{H} p_h(h) \left( \prod_{i} p_{z_i}(z_i | h) \right) = \sum_{H} \theta_h \left( \prod_{i} f_{\alpha}(z_i | \alpha_h) \right)$$

where $H$ is the number of F0 candidates, and $f_{\alpha}(z_i | \alpha_h)$ is the pdf of a Bernoulli distribution with success (i.e., speech presence) probability $\alpha$. What is important with this model is that each hidden variable $h$ is associated with an F0 candidate $\alpha_h$ and that the speech presence probabilities $(\alpha_h)_{h=1,\ldots,H}$ for each $h$ are designed so as to represent the harmonic structure for $\alpha_h$. Specifically, we define $\alpha_h$ by using a mixture of harmonically arranged generalized Gaussian distributions as (see Fig. 2)

$$\alpha_h = \beta_h \sum_m f_\rho(\xi_m; \mu, \rho, \lambda)$$

where $f_\rho(\xi_m; \mu, \rho, \lambda)$ is the pdf of a generalized Gaussian distribution with location $\mu$, scale $\rho$, and shape $\lambda$. In our current implementation, $\gamma = 30, \delta = 8, and we consider 37 F0 candidates (i.e., $H = 37$) between 85–255 Hz spaced by a semitone interval. We determine the scaling factor $\beta_h$ so that the maximum of $(\alpha_h)_{h=1,\ldots,H}$ becomes 1 $-$ $\epsilon$, where we use $\epsilon = 0.01$.

Finally, the prior distribution $p_h(h) = \theta_h$ of an F0 candidate is modeled by using a first Markov chain to make the contour of an F0 estimate smooth. The detailed is omitted due to limited space.

3.2. Inference

Now, we turn our attention to the estimation of clean speech given noisy speech. The first goal is to obtain the posterior distribution $\hat{z} = p_{z|x}(z | y)$ of a spectral mask. This posterior distribution is used to make a time-domain spectral masking filter $g_{\text{env}}(t) = g_{\text{IDCT}}(\hat{z})$, where $\hat{z} = [\hat{z}_1, \ldots, \hat{z}_T]$ and $t$ is the time-domain sample index.

To this end, we first derive some preliminary pdf’s. The conditional distribution of $y_i$ provided $z_i$ is given by

$$p_{y_{1:k}|z_i = 1, k} = p_{y_{1:k}}(y_i | k), \quad p_{y_{1:k}}(y_i | z_i = 0, k) = p_n(y_i),$$

where $p_{y_{1:k}}$ is given by (7). Then, we have the joint distribution of $y_i$ and $z_i$ given $k$ and $h$ as

$$p_{y_{1:k},h}(y_i, z_i | k, h) = p_{y_{1:k},h}(y_i | z_i, k) p_{z_i}(z_i | h),$$

where $p_{z_i}$ is given by a Bernoulli distribution (see (11)). The marginal distribution of $y_i$ given $k$ and $h$ is also easily derived as

$$p_{y_{1:k}}(y_i | k, h) = \sum_{z_i} p_{y_{1:k},h}(y_i, z_i | k, h).$$

The posterior distribution of a spectral mask is now easily obtained in a weighted sum form as

$$p_{z|x}(z_i = 1 | y) = \sum_k p_{k|x}(k | y) p_{z|x,h}(z_i = 1 | y, k, h)$$

where the numerator and denominator of (17) are given by (14) and (15), respectively, and $p_{k|x}(y_i | k, h)$ appearing in (18) is the product of the marginal distributions of $y_i$ given by (15) for all $l$ values.

The second goal is to obtain a time-domain filter $g_{\text{env}}(t)$ for spectral envelope enhancement. Speech enhancement is performed by applying the convolution of $g_{\text{env}}(t)$ and $g_{\text{noise}}(t)$ to the noisy speech signal. One way for estimating the spectral envelope enhancement filter is to estimate the log-PSD envelope in the log spectrum domain and construct a time-domain filter based on the estimated log spectrum envelope via IDCT. Instead, in this paper, we estimate it based on log mel-frequency feature enhancement since we also use a feature model as described in Section 4. The feature-based spectral envelope enhancement filter is calculated by using the conventional model-based feature enhancement approach (see Section 2).

4. COUPLING WITH FEATURE ENHANCEMENT

We found that the method presented in Section 3 sometimes suffered from poor enhancement results. To avoid this, this section combines the method in Section 3 and log mel-frequency spectrum feature enhancement, which leads to SPEAC.

In the failure cases, the joint posterior probability $p_{k|x}(k, h | y)$ of $k$ and $h$ given by (18) seemed inappropriate. This inappropriate joint posterior probability led to the poor enhancement results because an enhanced speech signal is calculated by using this joint posterior probability according to (16).

The cause of this problem is that a log spectrum has too many dimensions to represent a spectral envelope. It is important to note that the calculation of $p_{k|x}(k, h | y)$ is essentially a classification task. In general, the curse of dimensionality makes inference in a high dimensional space a tough task. This would account for the difficulty of achieving an appropriate calculation of $p_{k|x}(k, h | y)$.

With this in mind, SPEAC considers both a log spectrum $y$ and a log mel-frequency feature $g$ as being observed, thus enabling us to take advantage of the low-dimensional feature enhancement approach. The basic idea underlying the proposed domain coupling method is to consider a mixture component index $k$ as a junction of the log spectrum and feature domains (see Fig. 1 (c)). Specifically, we change the spectral envelope model (1) to the following model,
which generates both \( x \) and \( z \):

\[
p_{X,Z}(x, z) = \sum_k p_{X|Z}(x|k)p_{Z|X}(z|k)p_{K}(k),
\]

(19)

where we assume the conditional independence of \( x \) and \( z \) given \( k \).

Based on (19), we finally obtain an alternative to (18) as

\[
p_{K|Y}(k|h, y) = \frac{p_{Y|K}(y|h, k)p_{K}(k)}{\sum_k \sum_i p_{Y|K}(y|h, k)p_{K}(k)p_{K}(k)}.
\]

(20)

This means that we use a feature-domain posterior distribution \( p_{K}(k|y) \) instead of \( p_{K}(k) \) for a mixture component prior. This formulation also enables us to skip the evaluation of \( p_{K|Y}(k|y) \) for \( k \) with small \( p_{K}(k|y) \), which saves significant computational cost.

In addition, we perform voiced/unvoiced classification in the feature domain as in [6]. We use both the spectral masking filter \( \hat{g}_{\text{mask}}(t) \) and the spectral envelope enhancement filter \( \hat{g}_{\text{env}}(t) \) for voiced frames while only the latter is used for unvoiced frames.

### 4.1. Speech and noise model construction

Finally, we briefly describe how to construct a noise model \( p_{n}(n) \) and a combined speech model \( p_{X,Z}(x, z) \).

The proposed noise model construction method involves an initialization step and an update step. In the initialization step, we estimate the mean \( \mu_{n} \) and variance \( \sigma_{n}^{2} \) of the noise model from \( N = 20 \) observed frames at the beginning of an utterance. We perform speech enhancement once by using the initial noise model.

The update step compensates for the deviation of the actual noise from the initial noise model in each frame by using the posterior distribution of a spectral mask \( \hat{z}_{t} = p(z_{t} = 1|y) \) calculated in the first round enhancement. The mean is updated as \( \mu_{n}^{t} \leftarrow (N_{t-1} \mu_{n}^{t-1} + \hat{z}_{t})/(N + \hat{z}_{t}) \) while the variance is kept unchanged. Then, we perform speech enhancement again.

The combined speech model is constructed so that the log spectral envelope and log mel-frequency feature models share a mixture component index. To do this, we first train a feature GMM by using an EM algorithm. We save the sample weights obtained in the last E-step and use these weights to calculate the mean and variance of a log spectral envelope for each mixture component.

### 5. EXPERIMENTAL RESULTS

SPEAC was tested by using speech data contained in the Aurora2 corpus. The corpus consists of a training set and three test sets (A, B, and C). The training set consists of 8440 clean speech files, and we used it for speech model training. For evaluation, only test set A was used, where four different noise environments (subway, babble, car, exhibition) were considered. We used 20020 noisy speech files with SNR s of 0, 5, 10, 15, and 20 dBs. We chose OMLSA–IMCRA\(^1\) as a competitor.

Fig. 3 compares the segmental SNRs for observed noisy speech, enhanced speech obtained with OMLSA–IMCRA, and enhanced speech obtained with SPEAC. Each bar shows an average of segmental SNRs obtained for 5005 files (5 different input SNRs and 1001 files per SNR). SPEAC was superior or comparable to OMLSA–IMCRA for all of the four noise environments. In particular, with babble noise, the segmental SNR improvement offered by OMLSA–IMCRA was relatively limited. By contrast, SPEAC exhibited a significant segmental SNR improvement. This can be attributed to the use of a speech model and a harmonic structure, which helps us to distinguish target speech from babble in the background. For subway noise, which is very non-stationary, we also observed the great superiority of SPEAC. These results clearly show the effectiveness of SPEAC regardless of noise environments.

### 6. CONCLUSION

We have presented a speech enhancement method, called SPEAC, which restores both the spectral envelope and harmonic structure of noisy speech. SPEAC uses three models in combination: a harmonic structure model, a log spectral envelope model represented directly in the log spectrum domain, and a log mel-frequency feature model. The harmonic structure model is given in the form of a prior distribution of a spectral mask, thanks to which we can identify and remove noise components between harmonic peaks. The use of a feature model along with a log spectral envelope model makes it possible to combine the prior knowledge about the spectral envelope and the harmonic structure model while taking advantage of low dimensionality. Experimental results showed that SPEAC achieved a segmental SNR improvement that was comparable or superior to that of OMLSA–IMCRA.

### 7. REFERENCES


---

\(^1\)We used the implementation provided by the author of [1, 2], which is available at http://web.ee.technion.ac.il/People/IsraelCohen/Downloads/mlsma. We would like to thank the author for making it public.