Reduction of Highly Nonstationary Ambient Noise by Integrating Spectral and Locational Characteristics of Speech and Noise for Robust ASR

Tomohiro Nakatani, Shoko Araki, Marc Delcroix, Takuya Yoshioka, Masakiyo Fujimoto

NTT Communication Science Laboratories, NTT Corporation, Japan
nakatani.tomohiro@lab.ntt.co.jp

Abstract

This paper proposes a new multi-channel noise reduction approach that can appropriately handle highly nonstationary noise based on the spectral and locational features of speech and noise. We focus on a distant talking scenario, where a 2-ch microphone array receives a target speaker’s voice from the front while it receives highly nonstationary ambient noise from any direction. To cope well with this scenario, we introduce prior training not only for the spectral features of speech and noise but also for their locational features, and utilize them in a unified manner. The proposed method can distinguish rapid changes in speech and noise based mainly on their locational features, while it can reliably estimate the spectral shapes of the speech based largely on the spectral features. A filter-bank based implementation is also discussed to enable the proposed method to work in real time. Experiments using the PASCAL CHiME separation and recognition challenge task show the superiority of the proposed method as regards both speech quality and automatic speech recognition performance.

Index Terms: Noise reduction, nonstationary noise, automatic speech recognition, log power spectrum, direction of arrival

1. Introduction

When we capture our daily speech using distant microphones, various types of ambient noise, including rapidly time-varying noise, are mixed in the captured signals, and cause severe degradation of speech audible quality and automatic speech recognition (ASR).

Model based approaches have been extensively studied for overcoming this problem [1, 2]. With these approaches, the statistical characteristics of speech signals are often modeled by Gaussian mixture models (GMM) and/or hidden Markov models (HMM) based on prior training, and utilized to estimate and reduce the noise in the observation, for example, using the vector Taylor series (VTS) approximation technique [1]. It has been experimentally shown that stationary or relatively slowly time-varying noise may reliably be reduced by these approaches, which substantially improves the ASR performance. However, these approaches cannot accurately estimate rapidly time-varying noise due to the difficulty of noise tracking, and thus the improvement in the ASR performance is very limited.

In contrast, a number of attempts have been made to enhance the desired speech based on its location features using microphone arrays. One promising approach is based on binary/soft mask estimation (e.g., [3, 4]), where signals are assumed to be sparse so that each time-frequency (TF) bin is dominated only by speech or noise. Speech enhancement is realized by picking up TF bins whose location features correspond to the location of the desired signal. Because these approaches rely solely on the location features, they can effectively reduce the ambient noise no matter how rapidly the noise level changes, as long as their location features are different from that of the desired signal. However, because these approaches do not take any spectral characteristics into account, the quality of the separated signals is not necessarily good enough for certain speech applications, including ASR.

The goal of this paper is to propose a new noise reduction method that can simultaneously take advantage of both of the above existing approaches. This new method is an extension of a method proposed for source separation in [5], referred to as DOMinance based Locational and Power-spectral eHaracteristics INtegration (DOLPHIN). With DOLPHIN, the location features of individual source signals are modeled by location vector models (LM) and utilized jointly with their spectral features modeled by spectral HMMs to define a combined optimization criterion. Rapid changes in the individual source signals can be effectively tracked by using their location features, while the spectral shapes of the desired signals can be reliably estimated based on the spectral HMMs. In addition, it was shown that the combination of HMM parameters for the individual source signals can be searched in a computationally efficient manner by the expectation maximization (EM) algorithm.

Because DOLPHIN was originally proposed for source separation in almost anechoic environments, this paper extends it so that it achieves ambient noise reduction in reverberant environments. We assume a distant talking scenario, where a 2-ch microphone array is used for capturing sound. The direct speech signals and their reverberation come from the front and all directions, respectively, while the ambient noise may come from any direction. To model such complex location features of the speech and noise, respectively, we introduce location vector mixture models (LMM) into DOLPHIN. An LMM is defined as a mixture model composed of a set of LMs, and can represent a more complex location feature distribution than a single LM. Prior training of the LMMs as well as of the spectral GMMs is assumed to be conducted in advance using training data to allow DOLPHIN to distinguish the desired speech and noise. This paper also presents a method for implementing DOLPHIN using a filter bank for real-time processing.

2. DOLPHIN for ambient noise reduction

Suppose \( x_{m,k} \) is a short time Fourier transform of a signal captured at a microphone \( m (=1,2) \) and at a frequency \( k (=1, \ldots, N_k) \). Because the noise reduction method proposed in this paper is applied to each time frame independently as in section 2.2.1, the time frame indices of all symbols are omitted hereafter. Then, the observed signal is modeled as \( x_{m,k} = \sum_l s_{m,k}^{(l)} \), where \( s_{m,k}^{(l)} \) for \( l = 1 \) and \( l = 2 \), respectively, are the speech and the ambient noise signals captured at the \( m \)-th microphone. In the following, \( l \) is used as the index of the two sources, the speech (\( l = 1 \)) and the noise (\( l = 2 \)).

DOLPHIN uses two types of observed features: one is level normalized 2-ch observed signals, denoted as \( D_k \), and the other

\[ \text{To be concise, this paper does not use spectral HMMs. See [5] for how to incorporate spectral HMMs into DOLPHIN.} \]
is the log power spectra of 1-ch signals obtained by applying delay-and-sum beamforming to the 2-ch observed signals to enhance the front signal, denoted as \( X_k \). Letting \( T \) and \( | \cdot | \), respectively, be the non-conjugate transpose of a vector and the Euclidean norm of a vector, the two features are defined as

\[
D_k = x_k / |x_k|, \quad X_k = \ln\left(\sum_m x_m,k^2\right).
\]

Because \( D_k \) contains information about difference between channels, including interchannel phase and level differences, it is referred to as a location feature.

DOLPHIN estimates \( S^{(k)}_l \) for all \( k \), based on the above features. For this purpose, DOLPHIN introduces a generative model of the observed features as illustrated in Fig. 1. The model is composed of two sub-models shown on the left and right sides of the figure, which correspond to two generative models for \( D_k \) and \( X_k \), respectively.

To integrate the two sub-models, DOLPHIN utilizes a dominant sparseness assumption \([3]\), we assume that the location feature \( D_k \) is equal to the \( d^{(k)}_l \) value of the dominant source at each frequency. Then, the posterior pdf of \( D_k \) given \( L_k \) can be rewritten as

\[
p(D_k|L_k = l) = p(d^{(k)}_l = D_k; \phi^{(k)}_l),
\]

where \( p(d^{(k)}_l; \phi^{(k)}_l) \) is a pdf of \( d^{(k)}_l \), and \( \phi^{(k)}_l \) is its model parameter.

The location vector model (LM) was proposed to represent the pdf of a location feature corresponding to a point source, and was shown to be very effective for solving the problem of underdetermined blind source separation \([4]\). However, the pdf of \( d^{(k)}_l \) in the assumed scenario is more complex than that for source separation, because the probabilistic uncertainty of \( d^{(k)}_l \) is derived not only from the reverberation effect but also from the change in the noise source locations. To be more precise, the location features of the speech are concentrated around the front while those of the noise may be widely distributed over all the directions. To model the pdf of such a complex location feature, we use a location vector mixture model (LMM). An LMM for the \( d^{(k)}_l \) value of each source \( l \) at each frequency \( k \) is defined as

\[
p(d^{(k)}_l; \phi^{(k)}_l) = \sum_r w^{(k)}_{r,l} p(d^{(k)}_l | r; \phi^{(k)}_l),
\]

where \( F(d; \xi, \eta) = \frac{1}{\pi} \exp \left(-|d-\xi|^2 / \eta^2\right) \) is the pdf of an LM with a mean location vector \( \xi \) and a variance \( \eta \). \( H \) is the conjugate transpose of a vector, \( r \) is a mixture component index for \( l \) at a frequency \( k \), and \( w^{(k)}_{r,l} \) is a mixture weight for \( r \). Note that we also assume that (4) holds at individual frequencies independently for computationally efficient optimization based on the EM algorithm.

In this paper, \( \phi^{(k)}_l \) is assumed to be fixed\(^2\) in advance based on prior training using multi-channel databases for speech and noise in this paper. The learning algorithm for LMMs given in \([4]\) can be used for the prior training.

2.2. Model parameter estimation

DOLPHIN estimates the spectral Gaussian index pairs \( q \) at each time frame while it deals with the DSIs, \( L_k \), as hidden variables, where the likelihood function is defined as

\[
\mathcal{L}(q) = \sum_{(L_k)} \mathcal{L}_k(q),
\]

where

\[
\mathcal{L}_k(q) = \prod_{l \in L_k} p(D_k, X_k, L_k | q) \prod_{l \in L_k} u(q) \left| u(q) \right|^2,
\]

\( \left| u(q) \right|^2 \) is the spectral Gaussian index pair, \( q = [q^{(1)}, q^{(2)}] \), the joint pdf of \( X_k \) and \( L_k \) is derived as

\[
p(X_k, L_k = l | q) = \beta^{(l)}_{q^{(1)}, l}(X_k) \int_{-\infty}^{X_k} q^{(1), l}(S) dS,
\]

where \( l \) indicates the non-dominant source index.

2.1. DOLPHIN sub-models for ambient noise reduction

2.1.1. Sub-model for spectral feature

First, we model the log power spectra of speech and noise, denoted by \( S^{(k)}_l \), by spectral GMMs. With a spectral GMM, the distribution of \( S^{(k)}_l \) for each source \( l \) is modeled as

\[
p(S^{(k)}_l; \psi^{(l)}_k) = \sum_{q} u^{(q)}_k p(S^{(k)}_l | q; \psi^{(l)}_k),
\]

where \( \mathcal{N}(S; \mu, \sigma) \) is a Gaussian probability density function (pdf) with a mean \( \mu \) and a variance \( \sigma \), \( q \) is a mixture components parameter of \( \mu^{(l)}_q, \sigma^{(l)}_q \), and \( u^{(q)}_k \) is a mixture component index for \( l \), referred to as a spectral Gaussian index, \( u^{(q)}_k \) is a mixture weight for \( q \), and \( \psi^{(l)}_k \) is a model parameters composed of \( \mu^{(l)}_q, \sigma^{(l)}_q \) and \( u^{(q)}_k \) for all \( q \) and \( k \). Hereafter, we use a representation \( \beta^{(l)}_{q,l} (S) = p(S | q, \psi^{(l)}_k) \) in (2) for brevity, and assume \( \psi^{(l)}_k \) to be fixed in advance by prior training using speech and noise databases.

Then, to model the relationship between the source signal \( S^{(k)}_l \) and the observed signal, \( X_k \), we adopt the log-max model \([6]\) because the use of this model allows us to achieve efficient optimization based on the EM algorithm as discussed in \([5]\).

The relationship is defined as \( X_k = \max_l S^{(k)}_l \). Then, given
where \( \{ \cdot \} \) represents a set of variables at all frequencies. Then, as in [5], the combinations of model parameters over different sources can be estimated by the EM algorithm in a computationally efficient manner. The processing flow can be derived as in [5], but we omit the derivation in this paper due to the limited space. The resultant processing flow is summarized below in section 2.2.1.

### 2.2.1. Processing flow

Let \( M_k^{(i)} = p(L_k = l | X_k, D_k, q) \) be the posterior pdf of \( L_k \) given the spectral and locational features for each \( k \) and \( q \). Hereafter, \( M_k^{(i)} \) is referred to as a DSI posterior. Then, the overall procedure by DOLPHIN can be summarized as follows:

1. Set the initial estimate of \( M_k^{(i)} \) based only on the locational features. To be more precise, \( M_k^{(i)} = Q_k^{(i)} \), where
   \[
   Q_k^{(i)} = \frac{p(D_k | L_k = l)}{\sum_l p(D_k | L_k = l')}
   \]
   Hereafter, \( Q_k^{(i)} \) is referred to as a normalized locational posterior.

2. Iterate the following until convergence is achieved, aiming at improving the parameter estimation by integrating the locational and spectral characteristics.
   - (a) Update the spectral Gaussian index \( \hat{q}^{(i)} \) independently for each \( i \), so that it best fits the observation considering the DSI posterior (M-step).
     \[
     \hat{q}^{(i)} = \arg \max \Phi_i^{(i)}(q^{(i)})
     \]
     where \( \Phi_i^{(i)}(q) = M_k^{(i)} \log \beta_i^{(i)}(X_k^{(i)}) + (1 - M_k^{(i)}) \times \log \int_{-\infty}^{\infty} \rho_i^{(i)}(S)dS + \log \varphi_i^{(i)}(q^{(i)}) \) is the spectral matching function using the DSI posterior.
   - (b) Update the DSI posterior estimate \( M_k^{(i)} \) for each \( k \) and \( l \) considering both the spectral and locational features (E-step).
     \[
     M_k^{(i)} = \frac{Q_k^{(i)}(X_k, L_k = l | \hat{q}^{(i)})}{\sum_l Q_k^{(i)}(X_k, L_k = l' | \hat{q}^{(i)})}
     \]

3. Estimate \( S_k^{(i)} \) for each frequency \( k \) based on the minimum mean square error estimation (MMSE).
   \[
   S_k^{(i)} = \hat{M}_k^{(i)} X_k + (1 - \hat{M}_k^{(i)}) \int_{-\infty}^{\infty} S_k^{(i)} \beta_i^{(i)}(S)dS
   \]
   \[
   \int_{-\infty}^{\infty} \varphi_i^{(i)}(\tilde{q}^{(i)}) \tilde{q}^{(i)}(S)dS
   \]
   The enhanced speech waveform is then calculated by using an inverse Fourier transform with the phase of the observed signal followed by the overlap-add synthesis.

### 3. Filter bank based implementation for real-time processing

Because the above processing flow can be executed frame by frame independently, it can be applied to real-time processing. One problem that increases the computational complexity is the use of high resolution log-power spectra (HS) with a high feature dimension as the spectral features. A less computationally complex alternative is the use of filter bank (FB) domain features with a low feature dimension for DOLPHIN. In the following, we refer to the FB domain implementation of DOLPHIN as DOLPHIN-FB and compare it with DOLPHIN in the HS domain, which we refer to as DOLPHIN-HS.

There are several ways to implement DOLPHIN-FB. In this paper, we present a short time Fourier transform based method, which is illustrated in Fig. 2. In this method, the spectral features \( X_k \) and the normalized locational posteriors \( Q_k^{(i)} \) are first extracted in the HS domain and transformed into the FB domain. The spectral features, \( X_k \), can be transformed to those in the FB domain, \( \hat{X}_k \), by employing the usual FB analysis used for ASR as in (7), where \( k \) is the frequency index in the FB domain and \( F_{k,l} \) is a coefficient of the filter bank transformation. On the other hand, we approximate the locational posterior in the FB domain, \( \hat{Q}_k^{(i)} \), by using the average of the posterior in the HS domain weighted by the power of the signal and the filter bank transformation coefficient as in (8).

\[
\hat{X}_k = \text{ln}(\sum_k F_{k,l} \exp(X_k)) \quad (7)
\]
\[
\hat{Q}_k^{(i)} = \sum_k F_{k,l} \exp(X_k) Q_k^{(i)} / \sum_k F_{k,l} \exp(X_k) \quad (8)
\]

Using these values, we can then execute processing step 2 shown in section 2.2.1 in the FB domain, assuming that we have spectral GMMs trained on the FB domain spectral features in advance. Next, we re-estimate the DSI posterior, \( M_k^{(i)} \), in the HS domain based on (6) using the spectral Gaussian index pairs \( q \) estimated in the FB domain. For this purpose, we also assume we have spectral GMMs that are trained on the HS domain spectral features and that share the Gaussian indices with the spectral GMMs in the FB domain. Finally, the speech signal is estimated based on the MMSE in the HS domain.

The real-time factor (RTF) of DOLPHIN-HS was 4.52 while that of DOLPHIN-FB was 0.69 when using MATLAB interpreter in our experiments. This means DOLPHIN-FB can greatly improve the computational efficiency of DOLPHIN-HS.

### 4. Experiments

To evaluate the noise reduction performance, we used the isolated utterances of the development set in the PASCAL CHiME speech separation and recognition challenge [7]. We compared DOLPHIN-HS and DOLPHIN-FB with the VTS approach [1] (VTS) and a noise reduction method based only on LMMs in the HS domain (LMM). LMM is identical to DOLPHIN-HS except that it does not use any spectral models. By this comparison, we can test the effectiveness of the integration of spectral and locational characteristics for DOLPHIN. The mixture component numbers of the spectral GMMs and the LMMs were set, respectively, at 256 and 4 for both DOLPHIN-FB and DOLPHIN-HS. For speech, the speaker dependent spectral GMMs were trained on individual speakers in the training set, and the noise spectral GMMs were trained on all the noise data in the training set. The LMMs were trained using the same training set except that the LMMs for speech were speaker independent models and shared by all the speakers. The frame size and shift were set at 100 ms and 25 ms, respectively. The dimensions of the features, or the numbers of frequency bands, were set at 801 and 40 for
DOLPHIN-HS and DOLPHIN-FB, respectively. The EM iteration number was fixed at 5 in all the experiments.

4.1. Quality of enhanced speech

Figure 3 shows the quality of the enhanced speech depending on the signal-to-noise ratio (SNR) of the observation in terms of the average cepstral distortion (CD) calculated over the 1st to 12th order cepstral coefficients and the average segmental SNR. Segmental SNRs were calculated by first extracting noise remaining in the enhanced speech by subtracting the clean speech from the enhanced speech, and then by calculating the power ratios of the enhanced speech to the extracted noise.

According to the figure, under all SNR conditions, DOLPHIN-FB and DOLPHIN-HS substantially reduced the CDs and increased the segmental SNRs, and greatly outperformed VTS and LMM. By comparing DOLPHIN-HS and DOLPHIN-FB, we can confirm that DOLPHIN-FB was almost comparable to DOLPHIN-HS.

4.2. PASCAL CHiME keyword recognition task

We also evaluated the quality of the enhanced speech in terms of ASR performance. For the evaluation, we used the keyword recognition task defined for the PASCAL CHiME challenge [7]. Acoustic models trained on clean speech (clean-condition training) and on enhanced speech (multi-condition training) were used for the recognition. Multi-condition training data were artificially created by adding noise from the training samples to the clean speech training data at different SNRs. We used speaker dependent acoustic models consisting of left-to-right HMMs trained with the SOLON recognizer [8]. For the clean acoustic model, the total number of HMM states was 254 and each state had 7 Gaussians. For the multi-condition model, each HMM state had 20 Gaussians. Under this condition, the recognition accuracy of the ‘clean’ speech was 96.8%.

Table 1 shows the keyword recognition accuracy of each method. Again, both of DOLPHIN-HS and DOLPHIN-FB greatly outperformed the other methods and DOLPHIN-FB was almost comparable to DOLPHIN-HS except that slightly larger degradation can be confirmed with DOLPHIN-FB under low SNR conditions.

5. Summary

This paper proposed a new noise reduction method that can appropriately handle highly nonstationary noise by exploiting the spectral and locational features of speech and noise using microphone arrays. Two sets of spectral and locational models with prior training were introduced and integrated for efficiently distinguishing the speech and the noise. A technique to implement the proposed method in the filter bank domain was also presented aiming at realizing real-time processing. In the experiments, the proposed method greatly improved the quality of speech under various noise conditions in terms of cepstral distortions, segmental SNRs, and keyword recognition accuracies using the PASCAL CHiME challenge database. The results showed the effectiveness of the integration of spectral and locational features for noise reduction. The use of filter bank implementation was also effective to greatly reduce the computational complexity of the proposed method without much performance degradation.

6. References