

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

Phase-aware Speech Enhancement with Complex Wiener Filter

HUY NGUYEN, TUAN VU HO, MASATO AKAGI, (Life Member, IEEE), and MASASHI UNOKI, (Member, IEEE)

Graduate School of Advanced Science and Technology, Japan Advanced Institute of Science and Technology (JAIST), Nomi, Ishikawa 923-1292, Japan

Corresponding author: Huy Nguyen (hnguyen@jaist.ac.jp).

This work was supported by the Fund for the Promotion of Joint International Research (Fostering Joint International Research (B)(20KK0233) and by Grant-in-Aid for Scientific Research (B) (21H03463).

ABSTRACT In speech enhancement, accurate phase reconstruction can significantly improve speech quality. While phase-aware speech enhancement methods using the complex ideal ratio mask (cIRM) have shown promise, the estimation difficulty of the phase is shared with the real and imaginary parts of the cIRM. The pattern lacking in the imaginary part poses particular difficulties. To address this issue, we proposed a phase-aware speech enhancement method that uses a complex Wiener filter, which delegates the estimation of speech and noise amplitude properties and the phase property to different models, mitigating the issues with the cIRM and improving the effectiveness of neural-network training. Our method uses a speech-variance estimation model with a noise-robust vector-quantized variational autoencoder and a phase corrector that maximizes the scale-invariant signal-to-noise ratio in the time domain. To further improve speech-variance estimation, we propose a loss function that uses a categorical distribution of fundamental frequency (F0) for enhancing the spectral fine structure of estimated speech variance. We evaluated our method on the open dataset released by Valentini *et al.* to directly compare it with other speech-enhancement methods. Our method achieved a perceptual evaluation of speech quality score of 2.86 and short-time objective intelligibility score of 0.94, better than the state-of-the-art method based on cIRM estimation during the 2020 Deep Noise Challenge. Our comprehensive analysis shows that incorporating the proposed loss function for spectral-fine-structure enhancement improves speech quality, especially when the F0 is low.

INDEX TERMS Speech enhancement, complex Wiener filter, vector-quantized variational autoencoder, noise reduction, spectral fine structure enhancement, F0 distribution

I. INTRODUCTION

Speech quality and intelligibility deteriorate due to additive background noise. Various speech enhancement methods have been proposed to improve speech quality and intelligibility under such conditions. From traditional methods such as those using spectral subtraction [1], the Wiener filter [2], minimum-mean-squared-error estimation [3], and sub-space methods [4], which can only handle stationary noise, speech enhancement methods have evolved to handle other types of noise by incorporating deep neural networks.

There are mainly two processing domains with speech enhancement methods: time and time-frequency. Most speech enhancement methods in the time domain are based on powerful generative models, such as generative adversarial networks or diffusion probabilistic models, which attempt to generate a clean waveform directly, given the noisy speech

waveform or features [5]–[7]. Methods in the time-frequency domain operate on the time-frequency representation of the signal, which is obtained by short-time Fourier transform (STFT) or wavelet transform. A time-frequency representation is usually complex-valued and can generally be decomposed into amplitude and phase features. Most initial methods can only enhance the amplitude features while leaving the phase unprocessed [8] because the phase features have a complicated pattern and are difficult to model. The estimation targets are either the amplitude features or masks to modify the spectra, such as the ideal binary mask [9] or ideal ratio mask [10]. By incorporating deep neural networks, several speech enhancement methods with this approach perform excellently [11]–[13].

Speech signals follow a speech-production process and include linguistic, paralinguistic, and nonlinguistic information

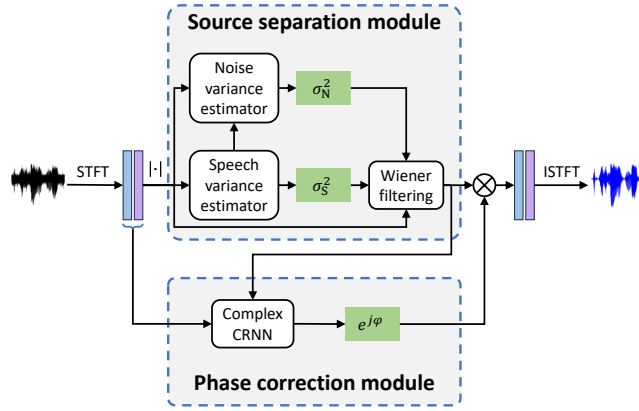


FIGURE 1. Overview diagram of the proposed speech enhancement model.

[14], while noise signals can be arbitrary. Thus, several studies have utilized the distribution of speech [15]–[17]. These studies modeled the speech and noise variances separately and then constructed the ideal ratio mask on the basis of Wiener filtering to enhance speech, in which a variational autoencoder (VAE) [18] modeled the distribution of speech and the non-negative matrix factorization [19] modeled the distribution of noise. With this approach, the generative speech model must satisfy noise-robustness and high-fidelity requirements to obtain high-quality enhanced speech.

In the time-frequency processing domain, besides the amplitude features, the importance of the phase features, with which speech quality improves significantly when the phase features are accurately estimated, has been clarified [20]. Therefore, several studies developed phase-aware enhancement methods [21]–[26], of which the most successful is based on the concept of the complex ideal ratio mask (cIRM) [27]. With the development of complex-valued deep neural networks [28], estimating the cIRM from the complex-valued noisy time-frequency representation [23]–[26] becomes possible. However, the unbounded property of the cIRM makes it difficult for optimization due to the infinite search space [29]. The imaginary part of the cIRM also lacks learnable patterns for the neural network to explore [30].

We aim to investigate a phase-aware speech enhancement approach that separately models speech amplitude, noise amplitude, and phase features to address the above issues with the cIRM. In our previously accepted conference paper [31], we introduced the complex Wiener filter for speech enhancement. The complex Wiener filter can be constructed using three parameters: speech variance, noise variance, and phase corrector. The process diagram of our method is illustrated in Figure 1. Amplitude spectrogram enhancement is delegated to the speech and noise variance estimation, while the phase corrector reconstructs the distorted phase spectrogram. Thus, the complex Wiener filter should require less estimation effort than the cIRM. We utilized a vector-quantized variational autoencoder (VQVAE) [32] with Itakura-Saito (IS) diver-

gence to capture the distribution of speech variance via a discrete latent space, which enables us to model high-quality and noise-robust speech variance. We used a complex-valued deep neural network to learn the phase corrector that maximizes the scale-invariant signal-to-distortion ratio (SISDR) in the time domain. However, in our previous study, the mechanism we used for speech variance estimation tended to approximate the spectral envelope information. Therefore, the spectral fine structure in the estimated speech variance is either weak or inaccurate, especially when the fundamental frequency (F0) is low. In addition, previous work lacked a comprehensive analysis and discussion regarding the efficacy of using a VQVAE or phase corrector.

This paper is an extension of our conference paper [31]. To overcome the remaining issues with our previous method, we propose a novel loss function that uses a categorical distribution of F0 to enable the speech variance model to capture the spectral fine structure even when F0 is low. Using the dataset released by Valentini *et al.* [33], we evaluated the overall performance of the proposed method and comprehensively analyzed the effectiveness of its components, i.e., the VQVAE for speech variance estimation, IS divergence, and the loss function for spectral-fine-structure enhancement, and the phase corrector.

In Section II, we introduce the statistical assumptions of speech, noise, and noisy speech signals in the complex time-frequency domain. In Section III, we present our speech enhancement method. We then describe our loss function for spectral-fine-structure enhancement in Section IV. We conducted objective evaluations to evaluate the performance of the proposed method, the experimental setup and results of which we present in Section V. We conclude the paper in Section VI with a summary.

II. STATISTICAL ASSUMPTIONS

For additive noise, in the short-time Fourier transform (STFT) domain, the noisy complex spectrogram $\mathbf{Y} \in \mathbb{C}^{F \times T}$ is the sum of the clean-speech complex spectrogram $\mathbf{S} \in \mathbb{C}^{F \times T}$ and the noise complex spectrogram $\mathbf{N} \in \mathbb{C}^{F \times T}$, where F is the number of frequency bins and T is the number of time frames. Let y_{ft} , s_{ft} , and n_{ft} represent the coefficients of $\mathbf{Y}$, $\mathbf{S}$, and $\mathbf{N}$, respectively, at a frequency bin index $f \in [0, F)$ and a frame index $t \in [0, T)$. The relation between y_{ft} , s_{ft} , and n_{ft} are described by the following equation

$$y_{ft} = s_{ft} + n_{ft}. \quad (1)$$

Also, we assume that the complex coefficients of $\mathbf{S}$ and $\mathbf{N}$ follow the circularly symmetric complex normal distribution; in other words,

$$s_{ft} \sim \mathcal{N}_{\mathbb{C}}(0, \sigma_{S,ft}^2), \quad (2)$$

$$n_{ft} \sim \mathcal{N}_{\mathbb{C}}(0, \sigma_{N,ft}^2), \quad (3)$$

where $\sigma_{S,ft}^2$ and $\sigma_{N,ft}^2$ represent the variances of speech and noise, respectively. Assuming that speech and noise are uncorrelated, the spectrogram coefficients of the noisy signal

then also follow the complex normal distribution of which variance is the sum of the speech and noise variances:

$$y_{ft} \sim \mathcal{N}_{\mathbb{C}}(0, \sigma_{S,ft}^2 + \sigma_{N,ft}^2). \quad (4)$$

Under such constraints, the power spectrograms of speech, noise, and mixture signals follow the exponential distribution as follows

$$|s_{ft}|^2 \sim \text{Exp}(\sigma_{S,ft}^2), \quad (5)$$

$$|n_{ft}|^2 \sim \text{Exp}(\sigma_{N,ft}^2), \quad (6)$$

$$|y_{ft}|^2 \sim \text{Exp}(\sigma_{S,ft}^2 + \sigma_{N,ft}^2), \quad (7)$$

where $\text{Exp}(\lambda)$ indicates the exponential distribution with the mean parameter $\lambda > 0$. The log-likelihood function for a sample v to belong to an exponential distribution $\text{Exp}(\lambda)$ is

$$L(\lambda|v) = -\frac{v}{\lambda} - \log \lambda, \quad (8)$$

$$= -\text{d}_{\text{IS}}(v \| \lambda) + \text{const}, \quad (9)$$

where d_{IS} is the Itakura-Saito (IS) divergence defined as

$$\text{d}_{\text{IS}}(v \| \lambda) = \frac{v}{\lambda} - \log \frac{v}{\lambda} - 1. \quad (10)$$

For the rest of the paper, since we use v and λ as the power spectrogram coefficient and its variance parameter for a time-frequency bin, we conveniently define the notation

$$\text{D}_{\text{IS}}(\mathbf{X} \| \boldsymbol{\lambda}) = \sum_{f,t} \text{d}_{\text{IS}}(|x_{ft}|^2 \| \lambda_{ft}), \quad (11)$$

where $\mathbf{X} \in \mathbb{C}^{F \times T}$ is a complex spectrogram of which element-wise variances are defined in $\boldsymbol{\lambda} \in \mathbb{R}_+^{F \times T}$.

III. SPEECH ENHANCEMENT USING COMPLEX WIENER FILTER

A. COMPLEX WIENER FILTER

Under the assumptions in Section II, if the speech and noise variances are known, the magnitude spectrogram of clean speech can be estimated by applying the ideal ratio mask, of which equation is defined via the Wiener filter as

$$|\hat{s}_{ft}| = |y_{ft}| \frac{\sigma_{S,ft}^2}{\sigma_{S,ft}^2 + \sigma_{N,ft}^2}. \quad (12)$$

This technique leaves the phase spectrogram unmodified, affecting speech quality and intelligibility in the low SNR conditions. To overcome this issue, we propose the complex Wiener filter by introducing a phase corrector term $e^{j\varphi_{ft}}$ to modify the complex spectrogram as follows

$$\hat{s}_{ft} = y_{ft} \frac{\sigma_{S,ft}^2}{\sigma_{S,ft}^2 + \sigma_{N,ft}^2} e^{j\varphi_{ft}}. \quad (13)$$

From (13), there are three parameters required to derive for our proposed speech enhancement: speech variance $\sigma_{S,ft}^2$, noise variance $\sigma_{N,ft}^2$, and phase corrector $e^{j\varphi_{ft}}$. As the speech signals have a unique production process while the noise can be arbitrary, we estimate the noise variance based on predicted speech variance. Also, as the amplitude is more

dominant in speech quality and intelligibility, we estimate the phase corrector based on the amplitude-enhanced complex spectrogram using a complex-valued deep neural network. The overview diagram of the proposed method is shown in Figure 1, where the details of the noise variance estimator, speech variance estimator, and phase correction are described in the following subsections. The overview diagram of the training flow is shown in Figure 2.

B. SPEECH VARIANCE ESTIMATION USING VECTOR-QUANTIZED VARIATIONAL AUTOENCODER

1) Vector-quantized Variational Autoencoder

Vector-quantized variational autoencoder (VQVAE) is a framework for learning the probability distribution of a dataset. The VQVAE assumes that each observation $\mathbf{x}$ in a dataset $\mathcal{X}$ is stochastically generated from a latent variable $\mathbf{z} \in \mathbb{R}^D$ following the discrete uniform distribution on a set of pseudovectors in a codebook $\mathcal{C} = \{\mathbf{e}_k\}_{k=1}^K$ via a conditional distribution (decoder) $p_{\theta}(\mathbf{x}|\mathbf{z})$. The VQVAE approximates the intractable posterior $p_{\theta}(\mathbf{z}|\mathbf{x})$ as follows

$$\mathbf{z}|\mathbf{x} \sim q_{\phi, \mathcal{C}}(\mathbf{z}|\mathbf{x}) = \delta(\mathbf{e}^*), \quad (14)$$

where $\delta(\cdot)$ denotes the Dirac delta distribution and $\mathbf{e}^*$ is determined by an encoding-quantization process as follows

$$\boldsymbol{\mu} \triangleq \boldsymbol{\mu}_{\phi}(\mathbf{x}) \quad (\text{encoding}), \quad (15)$$

$$\mathbf{e}^* = \arg \min_{\mathbf{e}_k \in \mathcal{C}} \|\boldsymbol{\mu} - \mathbf{e}_k\|_2^2 \quad (\text{quantization}). \quad (16)$$

The settings of VQVAE resemble a communication system in which the encoder compacts the input vector $\mathbf{x}$ into a code $\boldsymbol{\mu}$, the quantization process maps $\boldsymbol{\mu}$ to the nearest code $\mathbf{e}^*$ in the codebook, and the decoder reconstructs the input. The parameters θ , ϕ , and $\mathcal{C}$ can be obtained by minimizing

$$\mathcal{L}_{\text{VQVAE}}(\theta, \phi, \mathcal{C}) = -\mathbb{E}_{\hat{\mathbf{z}} \sim q_{\phi, \mathcal{C}}} [\log p_{\theta}(\mathbf{x}|\hat{\mathbf{z}})] + \|\text{sg}(\boldsymbol{\mu}) - \mathbf{e}^*\|_2^2 + \beta \|\boldsymbol{\mu} - \text{sg}(\mathbf{e}^*)\|_2^2, \quad (17)$$

where $\text{sg}(\cdot)$ is the stop-gradient operator. As the sampling process of $\hat{\mathbf{z}}$ is not differentiable, the VQVAE uses the reparameterization trick as follows

$$\hat{\mathbf{z}} = \boldsymbol{\mu} + \text{sg}(\mathbf{e}^* - \boldsymbol{\mu}). \quad (18)$$

2) Method for achieving noise-robustness

The VQVAE captures the data distribution via a codebook; in other words, the domain of the latent variable is reduced to codebook $\mathcal{C}$ only instead of $\mathbb{R}^D$. Therefore, if a VQVAE captures the distribution of clean speech, it will only output clean speech samples regardless of input. From this property, we propose first pre-training the VQVAE on clean speech. Then, in the main training stage, using the pre-trained codebook, we train another noisy encoder to denoise speech in latent space and fine-tune the decoder to achieve noise-robustness.

For pre-training, we apply VQVAE to model the distribution of the speech complex spectrogram $\mathbf{S}$ as

$$s_{ft}|\hat{\mathbf{z}} \sim \mathcal{N}_{\mathbb{C}}(0, \hat{\sigma}_{\hat{\theta}, S, ft}^2(\hat{\mathbf{z}})), \quad (19)$$

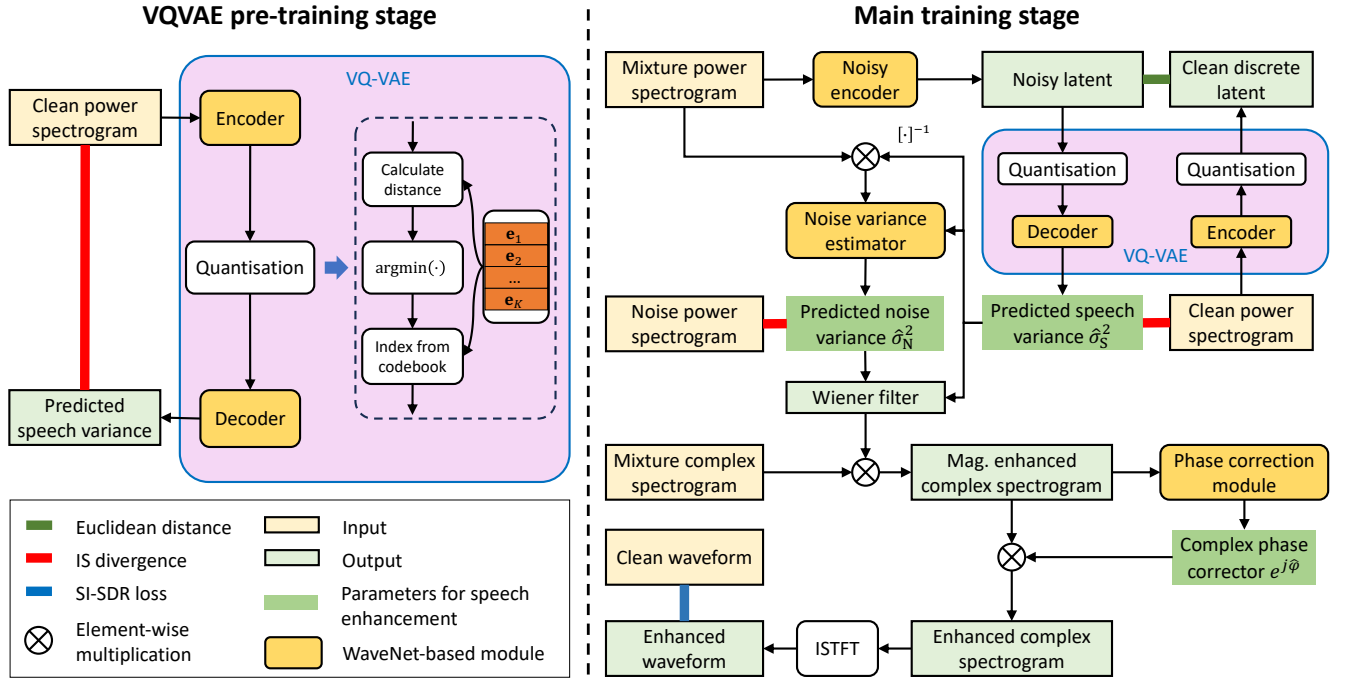


FIGURE 2. Block diagram of the two training stages.

which leads to the following loss function

$$\mathcal{L}_S(\theta, \phi_S, \mathcal{C}) = D_{\text{IS}}(\mathbf{S} \parallel \hat{\sigma}_S^2) + \|\text{sg}(\boldsymbol{\mu}) - \mathbf{e}^*\|_2^2 + \beta \|\boldsymbol{\mu} - \text{sg}(\mathbf{e}^*)\|_2^2. \quad (20)$$

where

$$\boldsymbol{\mu} \triangleq \boldsymbol{\mu}_{\phi_S}(\mathbf{S}) \quad (\text{encoding}), \quad (21)$$

$$\mathbf{e}^* = \arg \min_{\mathbf{e}_k \in \mathcal{C}} \|\boldsymbol{\mu} - \mathbf{e}_k\|_2^2 \quad (\text{quantization}), \quad (22)$$

$$\hat{\mathbf{z}} = \boldsymbol{\mu} + \text{sg}(\mathbf{e}^* - \boldsymbol{\mu}) \quad (\text{reparameterization}), \quad (23)$$

$$\hat{\sigma}_S^2 \triangleq \hat{\sigma}_{\theta, S}^2(\hat{\mathbf{z}}) \quad (\text{decoding}). \quad (24)$$

In the main training stage, the loss function for speech variance estimation is

$$\mathcal{L}_Y(\theta, \phi_Y) = D_{\text{IS}}(\mathbf{S} \parallel \hat{\sigma}_{S'}^2) + \beta \|\boldsymbol{\mu}' - \text{sg}(\mathbf{e}^*)\|_2^2, \quad (25)$$

where

$$\boldsymbol{\mu}' \triangleq \boldsymbol{\mu}_{\phi_Y}(\mathbf{S}) \quad (\text{noisy encoding}), \quad (26)$$

$$\mathbf{e}^{*'} = \arg \min_{\mathbf{e}_k \in \mathcal{C}} \|\boldsymbol{\mu}' - \mathbf{e}_k\|_2^2 \quad (\text{quantization}), \quad (27)$$

$$\hat{\mathbf{z}}' = \boldsymbol{\mu}' + \text{sg}(\mathbf{e}^{*'} - \boldsymbol{\mu}') \quad (\text{reparameterization}), \quad (28)$$

$$\hat{\sigma}_{S'}^2 \triangleq \hat{\sigma}_{\theta, S}^2(\hat{\mathbf{z}}') \quad (\text{decoding}). \quad (29)$$

C. NOISE VARIANCE ESTIMATION

The noise variance estimator is trained to reduce the IS divergence between the noise spectrogram and the predicted noise variance as follows

$$\mathcal{L}_N(\theta_N) = D_{\text{IS}}(\mathbf{N} \parallel \hat{\sigma}_{N, \theta_N}^2). \quad (30)$$

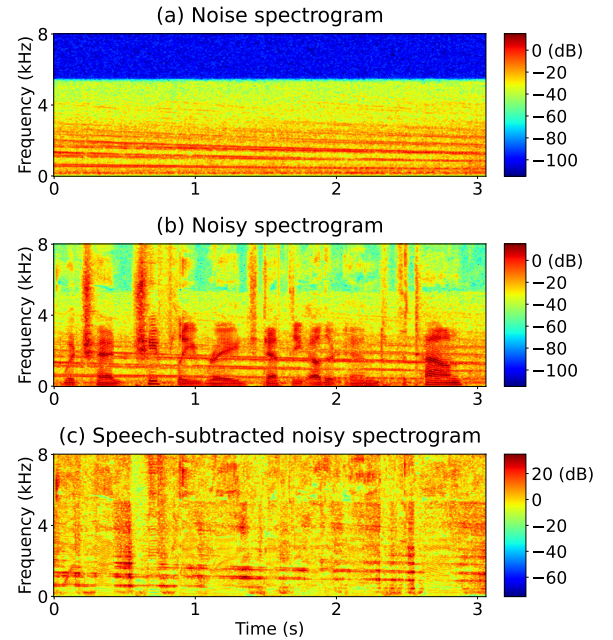


FIGURE 3. Visualization of (a) a noise log power spectrogram, (b) a noisy log power spectrogram, and (c) a noisy log power spectrogram subtracted by predicted log variance of speech. All the figures use the same noisy-clean pair of samples.

We empirically subtract the noisy speech log-power spectrogram from the estimated speech variance to condition the noise variance estimator on the estimated speech variance.

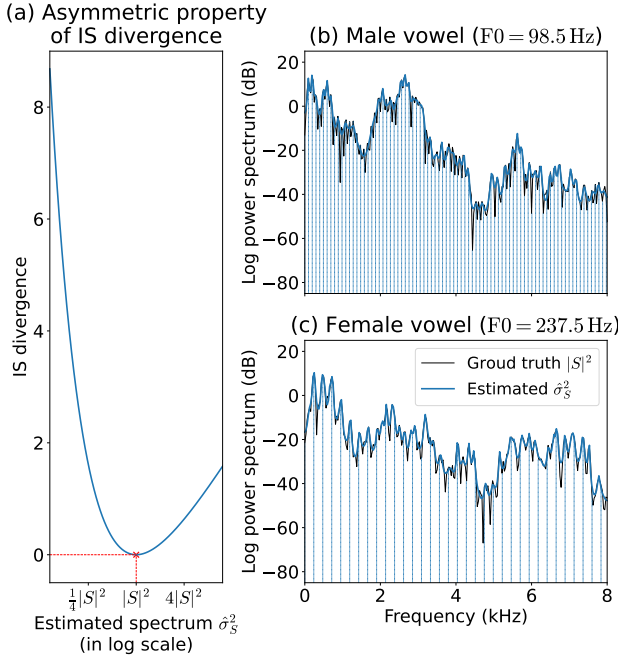


FIGURE 4. IS divergence and its envelope property in spectral estimation. Sub-figure (a) illustrates the asymmetric property of the IS divergence between a ground-truth power spectrum and an estimated power spectrum as a function of the estimated spectrum. The sub-figures (b) and (c) illustrate the low-rank spectrum estimation of the spectra of the vowel /æ/ pronounced by a male speaker (b) and a female speaker (c) by minimizing the IS divergence. The estimation model is an all-pole model with order $P = 128$, and the number of frequency bins is 257.

Although it is not entirely accurate, this results in a representation that resembles the noise log-power spectrogram better than the noisy log-power spectrogram where speech information is suppressed, as shown in Figure 3.

D. PHASE CORRECTION

Due to the phase warping property, it is challenging to model the distribution of the phase spectrogram directly. Several studies suggest using the scale-variant signal-to-noise ratio (SISDR) as the objective function to overcome this problem. The SISDR measures the difference between a ground-truth waveform y and an estimated waveform $\hat{y}$ as follows

$$\text{SISDR}(y, \hat{y}) = 10 \log_{10} \frac{\|y_{\text{target}}\|_2^2}{\|e_{\text{noise}}\|_2^2}, \quad (31)$$

$$y_{\text{target}} = \frac{\hat{y}^\top y}{\|\hat{y}\|_2} \cdot y, \quad (32)$$

$$e_{\text{noise}} = \hat{y} - y_{\text{target}}. \quad (33)$$

Therefore, to estimate the phase corrector term, we minimize the following objective

$$\mathcal{L}_P(\theta_P) = -\text{SISDR}(s, \hat{s}'), \quad (34)$$

$$s = \text{STFT}\{S\}, \quad (35)$$

$$\hat{s}' = \text{STFT}\{\hat{S} \odot e^{j\hat{\varphi}_{\theta_P}}\}, \quad (36)$$

where $\odot$ denotes Hadamard product and $\hat{S}$ is the magnitude-enhanced complex speech spectrogram. In other words, the phase corrector with parameters θ_P additively modifies the noisy phase to maximize the SISDR in the output waveform.

IV. SPECTRAL-FINE-STRUCTURE ENHANCEMENT BASED ON DISCRETE F0 DISTRIBUTION

Previous studies show that VQVAE with bottleneck tends to learn a highly discriminative codebook in linguistic information, while it can be invariant to the underlying speaker identity [34]. In addition, in the proposed model, the IS divergence in speech variance estimation also tends to estimate the spectral envelope of the spectrum [35] due to the asymmetric shape (see Figure 4). To further clarify this issue, we conduct a preliminary study using IS divergence to optimize an all-pole spectrum estimation model for spectra of the vowel /æ/ pronounced by a male and a female. The results illustrated in Figure 4 show that using IS divergence only estimates the spectral envelope and may omit the spectral fine structure when F0 is low. Therefore, we design a novel loss function to encourage the model to learn the spectral fine structure using the discrete F0 distribution of F0 candidates.

Given a log power spectrum $x \triangleq \{x(f)\}_f$, the significance [36] of an F0 candidate ξ can be defined as

$$q_\xi = \sum_{k \geq 1} \frac{x(k\xi) - x(k\xi - \xi/2)}{\sqrt{k}}. \quad (37)$$

The higher the significance q_ξ , the higher the chance that ξ is the F0 of $x(f)$. Therefore, given a set of C candidates $\{\xi_1, \dots, \xi_C\}$ for F0, we can estimate the F0 as the candidate with the highest significance. In other words, the F0 distribution is the Dirac delta distribution at the candidate ξ_{k^*} where $k^* = \arg \max_k q_{\xi_k}$. As the $\arg \max$ is not differentiable, we propose an F0 distribution approximation using the Softmax function with temperature as follows

$$p_{\xi_k} = \frac{\exp(q_{\xi_k}/\tau)}{\sum_{k'=1}^C \exp(q_{\xi_{k'}}/\tau)}, \quad (38)$$

where $\tau > 0$ is the temperature parameter. The lower the temperature, the closer the approximated distribution to the Dirac delta distribution, while the higher the temperature, the more uniform the approximated distribution [37].

Using the approximated categorical distribution of F0 as above, we can quantify the difference in F0 between two spectra via the Kullback-Leibler (KL) divergence between their F0 distributions. However, unvoiced or silent segments do not contain F0. Since there is no best F0 candidate, the approximated F0 distribution of an unvoiced segment is likely uniform, which means its entropy is high. Thus, we propose a loss function that quantifies the difference between a ground-truth spectrum x and an estimated spectrum $\hat{x}$ based on their F0 distributions $\mathbf{p} = [p_{\xi_1} \dots p_{\xi_C}]^\top$ and $\hat{\mathbf{p}} = [\hat{p}_{\xi_1} \dots \hat{p}_{\xi_C}]^\top$ as follows

$$\mathcal{L}_{F0} = \begin{cases} D_{\text{KL}}(\mathbf{p} \parallel \hat{\mathbf{p}}), & \text{if } H(\mathbf{p}) \leq \theta_h, \\ 0, & \text{otherwise} \end{cases} \quad (39)$$

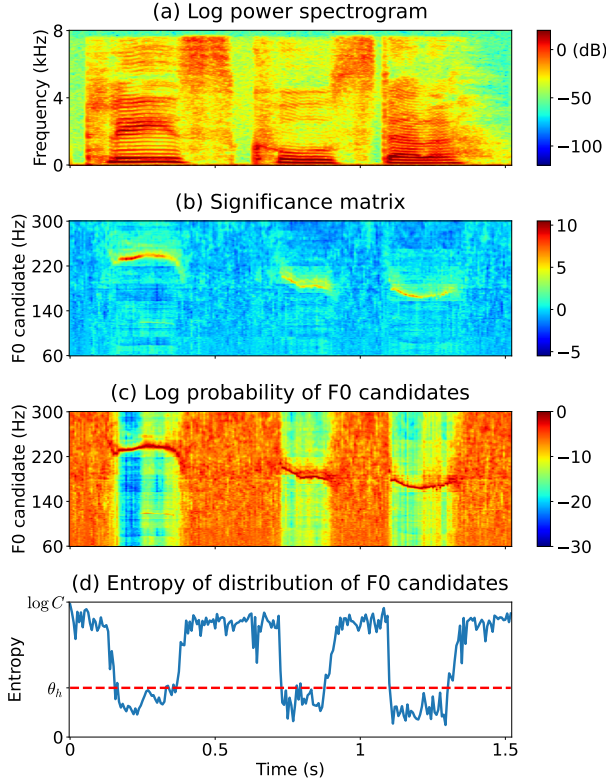


FIGURE 5. Visualization of outputs of each step in computing the discrete distribution of F0 candidates from the log power spectrogram of a speech signal: (a) the input log power spectrogram, (b) the significance matrix of which columns are significance vectors corresponding to the input, (c) the corresponding approximated log probability of F0 candidates, and (d) the corresponding entropy of these F0 distributions. The number of F0 candidates is $C = 241$, where the candidates are linearly located from 60 Hz to 300 Hz (resolution of 1 Hz).

where $\theta_h \in [0, \log C]$ is a fixed threshold. The KL divergence $D_{KL}(\mathbf{p} \parallel \hat{\mathbf{p}})$ from $\hat{\mathbf{p}}$ to $\mathbf{p}$ and the entropy $H(\mathbf{p})$ of $\mathbf{p}$ are respectively defined as

$$D_{KL}(\mathbf{p} \parallel \hat{\mathbf{p}}) = \sum_{k=1}^C p_{\xi_k} \log \frac{p_{\xi_k}}{\hat{p}_{\xi_k}}, \quad (40)$$

$$H(\mathbf{p}) = - \sum_{k=1}^C p_{\xi_k} \log p_{\xi_k}. \quad (41)$$

Figure 5 illustrates the significance matrix and the log probabilities and entropy of the approximated distributions of a speech spectrogram to visualize the quantities in the equations (37), (38), and (41) for each frame.

We apply the proposed spectral fine structure enhancement to our model by adding $\mathcal{L}_{F0}$ to the loss function of speech variance estimation in both pre-training and main training stages, in other words, in (20) and (25).

V. EXPERIMENTS

A. DATASET

We evaluate our proposed method using the open dataset from Valentini *et al.* [33]. This dataset has been used in several

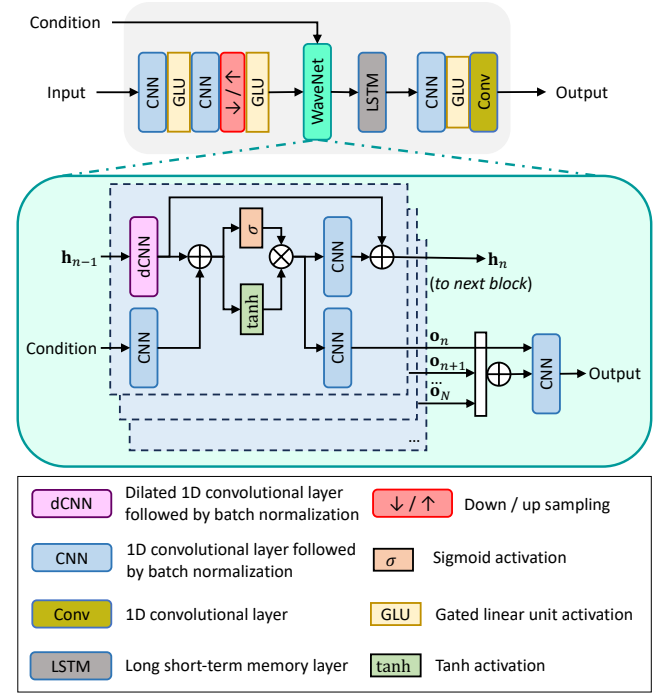


FIGURE 6. Architecture of WaveNet-based module.

recent speech enhancement studies, which we include here as baselines. The clean training set comprises 28 speakers (14 males and 14 females) and the test set of two speakers (one male and one female) from the Voice Bank corpus [38]. The noisy training set is constructed by mixing the clean training set with ten types of noise data at four signal-to-noise ratios (SNRs): 15 dB, 10 dB, 5 dB, and 0 dB. Eight of ten noise types are real recorded noise from the DEMAND dataset [39], while the other two (speech-shaped and babble) are artificially generated. The noisy test set is constructed by mixing the clean test set with five other types of noises from DEMAND dataset [39] at four SNRs: 17.5 dB, 12.5 dB, 7.5 dB, and 2.5 dB. The training and test sets have no speakers or noise types in common. All speech waveforms are resampled from 48 kHz to a 16 kHz sampling rate. To improve the robustness of the variation of input speech, we randomly scale the input speech between -35 dB and -20 dB.

B. MODEL AND TRAINING CONFIGURATIONS

In the implementation, the STFT uses the hanning window function with a window length of 25 ms (400 samples) and a hop length of 6.25 ms (100 samples). The number of points for the fast Fourier transform is 512, which results in 257 frequency bins.

The WaveNet-based module shown in Figure 6 is the basic block to construct our models. We use the VQVAE with a U-Net hierarchical structure of two layers (bottom and top) similar to [40] for the speech variance estimator. The bottom VQVAE aims to capture information in a higher temporal resolution, followed by the top, which tries to model the

features in a lower temporal resolution. The noise variance estimator and all the encoders and decoders in the speech variance estimator have the architecture of the WaveNet-based module without the long short-term memory (LSTM) layer. The inputs and outputs of the neural network are in log-compressed, e.g., log power spectrums and log variances. The phase correction module also follows the WaveNet-based module but with the LSTM layer, of which all the convolutional layers are complex-valued [28]. The output of the phase correction module is magnitude-normalized so that it only modifies the phase of the spectrogram.

For spectral-fine-structure enhancement, we use the F0 distribution over a set of $C = 241$ F0 candidates linearly located from 60 Hz to 300 Hz; in other words, the resolution is 1 Hz. To avoid numerical problems, we perform the mean-and-variance normalization on the significance matrix before applying the Softmax function. We set the temperature in the Softmax function to $\tau = 0.45$ and the entropy threshold to $\theta_h = 2$ (around 36% of the maximum entropy). These values are set to fit best the F0 detection in the training data.

The training procedure consists of two stages, shown in Figure 2, which are the pre-training and main-training stages. In the pre-training stage, the VQVAE is trained to estimate speech variance using the clean training set of 1000 epochs. We employ the data-dependent codebook re-estimation [41] and EMA algorithm [32] for codebook updating to obtain higher codebook perplexity. The noisy encoders, decoders, noise variance estimator, and phase correction network are trained for 1000 epochs in the main training stage. We use an Adam optimizer and a One-cycle Learning Rate scheduler with an initial learning rate of 5×10^{-4} and a maximum learning rate of 2×10^{-4} for all training stages.

C. EVALUATION METRICS

We use The Perceptual Evaluation of Speech Quality (PESQ) [42] and Short-Time Objective Intelligibility (STOI) [43] metrics to evaluate the overall performance of the proposed method. The PESQ scores, which range from -0.5 (bad) to 4.5 (excellent), measure the speech quality by comparing the enhanced signal to the clean reference speech signal. The STOI metric is highly correlated to perceptual speech intelligibility. The STOI scores range between 0 (lowest intelligibility) and 1 (highest intelligibility). For both metrics, a higher score indicates a better result.

D. OVERALL EVALUATION RESULTS

To ensure a fair comparison, we only compare the performance of our proposed model against other baseline models trained using the same Valentini *et al.* dataset [33]. From the results reported in Table 1, the proposed method outperforms many baselines and produces comparative results against other strong baselines, notably the state-of-the-art Deep Complex Convolution Recurrent Network (DCCRN) method [23] from the 2020 Deep Noise Suppression Challenge (DNS2020). Also, the proposed method improves the DCCRN+ model in Interspeech 2021 [26].

TABLE 1. Results of proposed and state-of-the-art methods trained on Valentini *et al.* dataset.

Method	PESQ-WB	STOI
Noisy	1.97	0.91
SEGAN, 2017 [5]	2.16	0.93
MMSE-GAN, 2018 [44]	2.53	0.93
Wave U-Net, 2018 [45]	2.40	–
MetricGAN, 2019 [46]	2.86	0.92
DCT-UNet, 2019 [47]	2.70	–
μ -law SGAN, 2020 [48]	2.86	0.94
DCCRN, 2020 [23]	2.68	–
DCCRN+, 2021 [26]	2.84	–
Proposed method	2.86	0.94

TABLE 2. Performance of the proposed method with different speech variance estimation model configurations on Valentini *et al.* dataset. All the results in this table are obtained without phase correction.

Speech variance models	PESQ-WB	STOI
VQVAE + d_{IS}	2.815	0.941
VQVAE + $d_{IS} + \mathcal{L}_{F0}$	2.817	0.942
VQVAE + d_{LSD}	2.689	0.940
VAE + d_{IS}	2.623	0.936

E. ABLATION STUDY

1) The Effectiveness of VQVAE in Speech Variance Modeling

In this section, we analyze the use of VQVAE for latent denoising in the speech variance estimation. For this purpose, we compare the speech variance estimation using VQVAE with one using a variational autoencoder (VAE), which Bando *et al.* [15] and Fang *et al.* [17] utilize in their speech enhancement methods. In VQVAE, the prior distribution of the latent variable is discrete uniform on the codebook $\mathcal{C}$ and the variational approximation of the posterior is modeled by the Dirac delta distribution as in (14). On the other hand, in VAE, the prior distribution is a standard Gaussian distribution, the posterior model is a Gaussian distribution as follows

$$\mathbf{z}|\mathbf{x} \sim q_{\phi}(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\mu_{\phi}(\mathbf{x}), \text{diag}(\sigma_{\phi}^2(\mathbf{x}))) . \quad (42)$$

For a fair comparison, we replace the VQVAE with the VAE while keeping the same latent dimension, network, and training configurations. In other words, the output dimensions of the VAE encoders are twice as large as the output dimensions of the VQVAE to model the mean and log variance of the posterior distribution. The reconstruction term in VAE is also IS divergence as in VQVAE, while the latent denoising is the KL divergence [17] instead of the quantization error as the second term in (25).

The results in the first and the fourth rows of Table 2 show that the VQVAE outperforms VAE as a speech variance estimation model for speech enhancement. Also, from the visualization in Figure 7, the VAE gave a less accurate and detailed speech variance. These results show the importance of the pre-training mechanism. In VQVAE, similar inputs are quantized to the same code in the codebook during the training, and this codebook is re-used in the denoising. On

TABLE 3. Improvement of PESQ-WB and STOI for each speaker in the test set of Valentini *et al.* dataset when applying $\mathcal{L}_{F0}$. All the results in this table are obtained without phase correction.

Speaker	PESQ-WB impr. (%)	STOI impr. (%)
p232 (male)	0.434	0.006
p257 (female)	0.340	0.144
Overall	0.385	0.078

TABLE 4. Performance of the model with and without phase correction on the test set of Valentini *et al.* dataset.

$\mathcal{L}_{F0}$	Phase correction	PESQ-WB	STOI
$\times$	$\times$	2.815	0.941
$\times$	$\checkmark$	2.857	0.943
$\checkmark$	$\times$	2.817	0.942
$\checkmark$	$\checkmark$	2.861	0.944

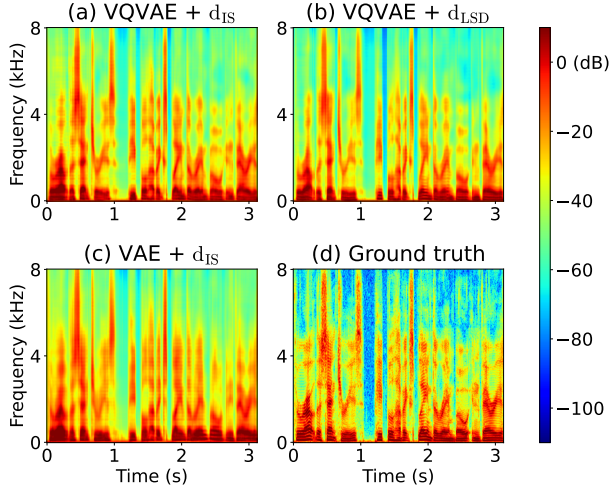


FIGURE 7. Example of the effect of different speech variance estimation model configurations. Sub-figures (a), (b), and (c) visualize the estimated speech log variances using the VQVAE with IS divergence, VQVAE with log spectral distance, and VAE with IS divergence. Sub-figure (d) is the ground-truth speech log spectrogram.

the other hand, in VAE, all the inputs are dragged to the same standard normal distribution, which is less discriminative. In other words, the denoising stage in VQVAE receives more benefits than in VAE, thanks to the codebook, which affirms our idea of using VQVAE for speech variance modeling.

2) The Role of IS Divergence

This section analyzes the use of IS divergence in the model. Although the IS divergence is necessary for the proposed speech enhancement method due to the assumption in Section II, it tends to ignore the spectral fine structure mentioned in Section IV. Thus, we investigate the effect of replacing the IS divergence with the log spectral distortion (LSD) - a common metric in the spectral domain, of which the equation is

$$d_{LSD}(x \parallel y) = \frac{1}{2} \left(\log \frac{x}{y} \right)^2. \quad (43)$$

The first and the third rows of Table 2 show that the model performance when using LSD is worse than when using IS divergence, even though the speech variance estimated by optimizing LSD has more similar detailed patterns. Using LSD changes the statistical assumption of speech and noise power spectrograms from the exponential distribution to the log-normal distribution. Under this assumption, the distribution of

the noisy speech also changes, where its parameter no longer has the additive property as in (7). Therefore, the Wiener filter is no longer valid. Thus, even with more detailed patterns, the LSD is not a solution to improve spectral fine structure due to assumption violation.

3) The Effectiveness of $\mathcal{L}_{F0}$

We analyze the effectiveness of the proposed spectral-fine-structure enhancement method by comparing the model's performance with and without $\mathcal{L}_{F0}$. The results reported in the first and the second rows of Table 2 show that using $\mathcal{L}_{F0}$ can slightly improve speech enhancement performance. For further analysis, we evaluate the improvement percentage of PESQ-WB and STOI for each speaker in the test set when applying $\mathcal{L}_{F0}$, where the improvement percentage of a metric after an operation is defined as

$$\text{Impr.}(\%) = \frac{\text{Final} - \text{Initial}}{\text{Initial}} \times 100\%, \quad (44)$$

where Final and Initial are metric values of the model before and after applying that operation.

The results in Table 3 show that the PESQ-WB and STOI improve overall, which means the proposed fine-structure enhancement technique helps model speech variance better. The improvement percentage of PESQ-WB is higher for the male speaker, which aligns with our original purpose when proposing $\mathcal{L}_{F0}$ to improve spectral fine structure when F0 is low. On the other hand, the improvement percentage of STOI is lower for the male speaker. Our hypothesis for the cause of this result is the limited capacity of the VQVAE. There is a trade-off when modeling the spectral envelope and spectral fine structure of highly fluctuated spectra from the male speaker. The spectral envelope is more associated with linguistic information and speech intelligibility than the spectral fine structure. Therefore, the fine structure enhancement does not improve STOI much for the male speaker. This trade-off is less severe for the female speaker with a higher F0 and less fluctuated spectral fine structure. Before using the proposed spectral enhancement, the estimated spectral fine structure is already better for higher F0 spectra, which results in a lower improvement in PESQ-WB. However, with the higher improvement of STOI for the female speaker, the $\mathcal{L}_{F0}$ forces the model to utilize its capacity in modeling speech variance.

Figure 8 illustrates the estimated speech variances and Wiener filters with and without using spectral-fine-structure enhancement for some samples in the Valentini *et al.* test

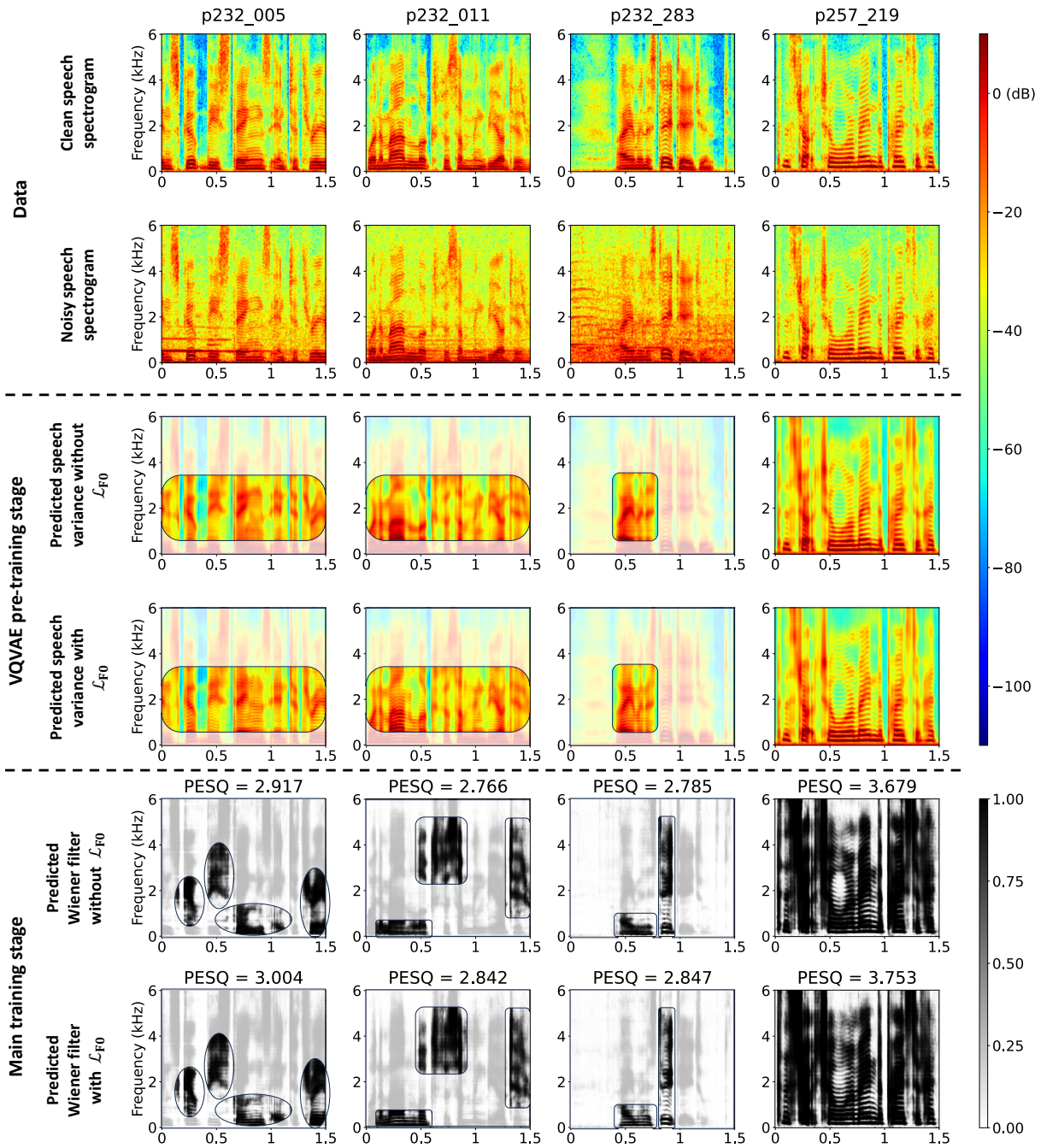


FIGURE 8. Samples of the effect of $\mathcal{L}_{F0}$ in the estimated variance of speech power spectrogram in pre-training and main training (denoising) stages on four different samples (four columns) from the Valentini test set. The top figures are the clean and noisy speech power spectrograms as references. The differences between the outputs of the proposed model with and without using $\mathcal{L}_{F0}$ are highlighted in the first three samples.

dataset [33]. The figure visualizes the phenomena we described above and confirms our hypothesis for the effectiveness of the proposed $\mathcal{L}_{F0}$. For the frames with low F0 in the first three samples, the use of $\mathcal{L}_{F0}$ improves the spectral fine structure of predicted speech variance significantly in the pre-training stage, which results in a Wiener filter estimation with closer detailed patterns to the clean speech spectrogram. Nevertheless, the formants and formant transitions are more

apparent when not using $\mathcal{L}_{F0}$, especially with the weak-harmonic frames. For the frames with high F0, the effect of $\mathcal{L}_{F0}$ is not visually significant.

4) The Effectiveness of Phase Correction

This section analyzes the effectiveness of the phase correction module in the proposed speech enhancement method. The results reported in Table 4 indicate that the phase correction

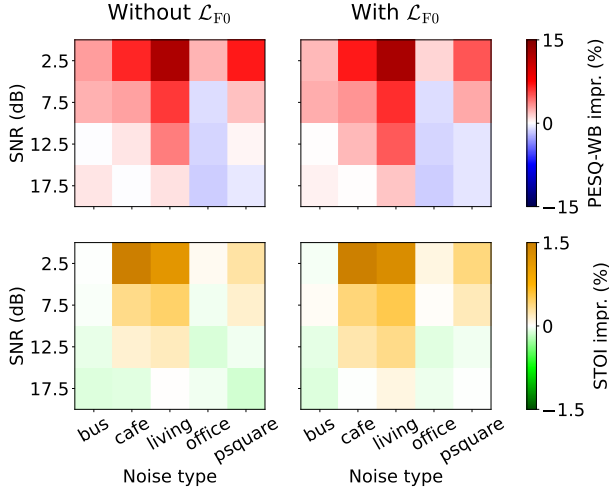


FIGURE 9. Effect of phase correction module in model performance. The left sub-figures show the improvement percentage in PESQ (top) and STOI (bottom) for our model without $\mathcal{L}_{F0}$, while the right sub-figures display the same metrics with the inclusion of $\mathcal{L}_{F0}$. There is little difference between left and right figures, so the phase correction consistently improves the metrics.

module can significantly improve the quality and intelligibility of enhanced speech. This improvement is consistent when incorporating with the spectral-fine-structure enhancement loss function $\mathcal{L}_{F0}$. Also, the phase difference between noisy speech and clean speech is strongly associated with the local signal-to-noise ratio (SNR) at a time-frequency bin [49] as follows

$$\angle y_{ft} = \angle s_{ft} + \angle \left\{ 1 + \frac{n_{ft}}{s_{ft}} \right\}, \quad (45)$$

where $\angle \{\cdot\}$ indicates the phase of a complex number. The local SNR depends on the noise characteristics, including noise level and type. Therefore, we evaluate the PESQ-WB and STOI improvement when applying the phase correction to different noise types and SNRs in the Valentini *et al.* test dataset [33]. The evaluation results are visualized in Figure 9. The results show that incorporating phase correction is more effective in low SNR conditions, improving the PESQ-WB to 15% and STOI to 1.5%. In contrast, phase correction can cause negative effects in high SNR conditions for some types of noise, which reduces the evaluation metrics. In low SNR conditions, the performance improves the most in living, cafe, and psquare noise types. Most noise signals in the living type contain music or singing voice. The noise signals in the cafe and psquare types are babble noise recorded in the crowded environment [33]. Such types of noise have formants in their spectra, where the power is concentrated at some frequencies. Thus, the local SNR is extremely low at those frequency bins. In contrast, the spectra of bus and office noises spread broader along the frequency axis, distributing the SNR uniformly in all frequency bins. Therefore, phase correction is essential to suppress noises in which power spectral density is concentrated.

VI. CONCLUSION

This paper evaluated our phase-aware speech enhancement method that uses a complex Wiener filter. The speech variance is estimated using a noise-robust VQVAE, and a phase corrector is approximated to maximize the SISDR of the waveform. We also proposed a loss function using a categorical distribution of F0 to further enhance the spectral fine structure of the estimated speech variance. Using PESQ-WB and STOI as objective evaluations on the Valentini *et al.* dataset, we comprehensively analyzed the effectiveness of the VQVAE, the phase corrector, and the loss function for spectral-fine-structure enhancement of our method.

The results indicate that our method achieved a PESQ-WB of 2.86 and STOI of 0.94, which are better than those of the state-of-the-art DCRNN model in the Deep Noise Suppression Challenge 2020, confirming the effectiveness of our method in enhancing speech quality and intelligibility. The VQVAE outperformed the vanilla VAE in speech variance estimation regarding both quality and noise-robustness. The proposed loss function for the spectral-fine-structure enhancement improved both the PESQ-WB and STOI of enhanced speech, especially for male speakers with lower F0. The improvement of the fine structure is visually observable in our sample illustrations. The phase corrector in the complex Wiener filter is crucial for speech enhancement, especially for high-level noise where the power is concentrated at certain frequencies instead of uniformly distributed. Our implementation source code and demo audios are available at <https://huynguyenqc.github.io/jekyll/update/2023/09/26/ieeeaccess-nhau23.html>.

REFERENCES

- [1] S. F. Boll, "Suppression of Acoustic Noise in Speech Using Spectral Subtraction," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 27, no. 2, pp. 113–120, 1979.
- [2] J. Chen, J. Benesty, Y. Huang, and S. Doclo, "New insights into the noise reduction wiener filter," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1218–1234, 2006.
- [3] Y. Ephraim and D. Malah, "Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, 1984.
- [4] K. Hermus, P. Wambacq, and H. V. hamme, "A review of signal subspace speech enhancement and its application to noise robust speech recognition," *EURASIP Journal on Advances in Signal Processing*, vol. 2007, no. 1, Sep. 2006.
- [5] S. Pascual, A. Bonafonte, and J. Serrà, "Segan: Speech enhancement generative adversarial network," in *Proceedings of Interspeech*, 2017.
- [6] Y. Luo and N. Mesgarani, "Conv-tasnet: Surpassing ideal time-frequency magnitude masking for speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 8, pp. 1256–1266, 2019.
- [7] Y.-J. Lu, Y. Tsao, and S. Watanabe, "A study on speech enhancement based on diffusion probabilistic model," in *2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2021, pp. 659–666.
- [8] Y. Xu, J. Du, L. R. Dai, and C. H. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 23, no. 1, pp. 7–19, 2015.
- [9] D. Wang, *On Ideal Binary Mask As the Computational Goal of Auditory Scene Analysis*. Boston, MA: Springer US, 2005, pp. 181–197.
- [10] C. Hummersone, T. Stokes, and T. Brookes, *On the Ideal Ratio Mask as the Goal of Computational Auditory Scene Analysis*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2014, pp. 349–368.

- [11] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising autoencoder," in *Proc. Interspeech 2013*, 2013, pp. 436–440.
- [12] J. Chen and D. Wang, *DNN Based Mask Estimation for Supervised Speech Separation*. Cham: Springer International Publishing, 2018, pp. 207–235.
- [13] S. Braun and I. Tashev, "Data augmentation and loss normalization for deep noise suppression," in *Speech and Computer: 22nd International Conference, SPECOM 2020, St. Petersburg, Russia, October 7–9, 2020, Proceedings*. Berlin, Heidelberg: Springer-Verlag, 2020, p. 79–86.
- [14] P. Roach, R. Stibbard, J. Osborne, S. Arnfield, and J. Setter, "Transcription of prosodic and paralinguistic features of emotional speech," *Journal of the International Phonetic Association*, vol. 28, no. 1-2, p. 83–94, 1998.
- [15] Y. Bando, M. Mimura, K. Itoyama, K. Yoshii, and T. Kawahara, "Statistical speech enhancement based on probabilistic integration of variational autoencoder and non-negative matrix factorization," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 716–720.
- [16] S. Leglaive, L. Girin, and R. Horaud, "A variance modeling framework based on variational autoencoders for speech enhancement," in *2018 IEEE 28th International Workshop on Machine Learning for Signal Processing (MLSP)*, 2018, pp. 1–6.
- [17] H. Fang, G. Carbajal, S. Wermter, and T. Gerkmann, "Variational autoencoder for speech enhancement with a noise-aware encoder," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 676–680.
- [18] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14–16, 2014, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2014.
- [19] C. Févotte, N. Bertin, and J.-L. Durrieu, "Nonnegative matrix factorization with the itakura-saito divergence: With application to music analysis," *Neural Computation*, vol. 21, no. 3, pp. 793–830, 2009.
- [20] K. Paliwal, K. Wójcicki, and B. Shannon, "The importance of phase in speech enhancement," *Speech Communication*, vol. 53, no. 4, pp. 465–494, 2011.
- [21] M. Strake, B. Defraene, K. Fluyt, W. Tirry, and T. Fingscheidt, "INTER-SPEECH 2020 Deep Noise Suppression Challenge: A Fully Convolutional Recurrent Network (FCRN) for Joint Dereverberation and Denoising," in *Proceedings of Interspeech*, 2020, pp. 2467–2471.
- [22] X. Li and R. Horaud, "Online monaural speech enhancement using delayed subband LSTM," in *Proceedings of Interspeech*, 2020, pp. 2462–2466.
- [23] Y. Hu, Y. Liu, S. Lv, M. Xing, S. Zhang, Y. Fu, J. Wu, B. Zhang, and L. Xie, "DCCRN: Deep complex convolution recurrent network for phase-aware speech enhancement," in *Proceedings of Interspeech*, 2020, pp. 2472–2476.
- [24] X. Hao, X. Su, R. Horaud, and X. Li, "Fullsubnet: A full-band and sub-band fusion model for real-time single-channel speech enhancement," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6633–6637.
- [25] N. L. Westhausen and B. T. Meyer, "Dual-signal transformation LSTM network for real-time noise suppression," in *Proceedings of Interspeech*, 2020, pp. 2477–2481.
- [26] S. Lv, Y. Hu, S. Zhang, and L. Xie, "DCCRN+: Channel-Wise Subband DCCRN with SNR Estimation for Speech Enhancement," in *Proc. Interspeech 2021*, 2021, pp. 2816–2820.
- [27] D. S. Williamson, Y. Wang, and D. Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 24, no. 3, pp. 483–492, 2015.
- [28] C. Trabelsi, O. Bilaniuk, Y. Zhang, D. Serdyuk, S. Subramanian, J. F. Santos, S. Mehri, N. Rostamzadeh, Y. Bengio, and C. J. Pal, "Deep complex networks," in *International Conference on Learning Representations*, 2018.
- [29] H.-S. Choi, J.-H. Kim, J. Huh, A. Kim, J.-W. Ha, and K. Lee, "Phase-aware speech enhancement with deep complex u-net," in *International Conference on Learning Representations*, 2018.
- [30] M. Hasannezhad, Z. Ouyang, W.-P. Zhu, and B. Champagne, "Speech enhancement with phase sensitive mask estimation using a novel hybrid neural network," *IEEE Open Journal of Signal Processing*, vol. 2, pp. 136–150, 2021.
- [31] T. V. Ho, Q. H. Nguyen, M. Akagi, and M. Unoki, "Vector-quantized Variational Autoencoder for Phase-aware Speech Enhancement," in *Proc. Interspeech 2022*, 2022, pp. 176–180.
- [32] A. van den Oord, O. Vinyals, and K. Kavukcuoglu, "Neural discrete representation learning," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017, pp. 6309–6318.
- [33] C. Valentini-Botinhao, X. Wang, S. Takaki, and J. Yamagishi, "Speech Enhancement for a Noise-Robust Text-to-Speech Synthesis System Using Deep Recurrent Neural Networks," in *Proc. Interspeech 2016*, 2016, pp. 352–356.
- [34] J. Chorowski, R. J. Weiss, S. Bengio, and A. van den Oord, "Unsupervised speech representation learning using wavenet autoencoders," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 12, pp. 2041–2053, 2019.
- [35] J. D. Markel and A. H. Gray, *Linear Prediction of Speech*. Springer Berlin Heidelberg, 1976, vol. 12.
- [36] T. Wang, W. Zhu, Y. Gao, J. Feng, and S. Zhang, "Hgc: Harmonic gated compensation network for speech enhancement," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 371–375.
- [37] E. Jang, S. Gu, and B. Poole, "Categorical reparameterization with gumbel-softmax," in *International Conference on Learning Representations*, 2017.
- [38] C. Veaux, J. Yamagishi, and S. King, "The voice bank corpus: Design, collection and data analysis of a large regional accent speech database," in *2013 International Conference Oriental COCOSDA held jointly with 2013 Conference on Asian Spoken Language Research and Evaluation (O-COSDA/CASLRE)*, 2013, pp. 1–4.
- [39] J. Thiemann, N. Ito, and E. Vincent, "The Diverse Environments Multichannel Acoustic Noise Database (DEMAND): A database of multichannel environmental noise recordings," *Proceedings of Meetings on Acoustics*, vol. 19, no. 1, p. 035081, 06 2013.
- [40] T. V. Ho and M. Akagi, "Non-parallel Voice Conversion based on Hierarchical Latent Embedding Vector Quantized Variational Autoencoder," in *Proc. Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge 2020*, 2020, pp. 140–144.
- [41] A. Lancucki, J. Chorowski, G. Sanchez, R. Marxer, N. Chen, H. J. Dolfing, S. Khurana, T. Alumae, and A. Laurent, "Robust training of vector quantized bottleneck models," *Proceedings of the International Joint Conference on Neural Networks*, 7 2020.
- [42] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221)*, vol. 2, 2001, pp. 749–752 vol.2.
- [43] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2010, pp. 4214–4217.
- [44] M. H. Soni, N. Shah, and H. A. Patil, "Time-frequency masking-based speech enhancement using generative adversarial network," *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5039–5043, 2018.
- [45] D. Stoller, S. Ewert, and S. Dixon, "Wave-u-net: A multi-scale neural network for end-to-end audio source separation," *ArXiv*, vol. abs/1806.03185, 2018.
- [46] S.-W. Fu, C.-F. Liao, Y. Tsao, and S.-D. Lin, "Metricgan: Generative adversarial networks based black-box metric scores optimization for speech enhancement," in *International Conference on Machine Learning*. PMLR, 2019, pp. 2031–2041.
- [47] C. Geng and L. Wang, "End-to-end speech enhancement based on discrete cosine transform," in *2020 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA)*. IEEE, 2020, pp. 379–383.
- [48] H. Li, Y. Xu, D. Ke, and K. Su, " μ -law sgan for generating spectra with more details in speech enhancement," *Neural Networks*, vol. 136, pp. 17–27, 2021.
- [49] N. Zheng and X.-L. Zhang, "Phase-aware speech enhancement based on deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 1, pp. 63–76, 2019.



HUY NGUYEN received the B.S. degree from Vietnam National University, Ho Chi Minh, in 2019, following by an M.S. degree in information science from the Japan Advanced Institute of Science and Technology (JAIST), in 2021. Currently, he is pursuing a Ph.D. degree at JAIST with the Unoki Laboratory. His research interests include speech enhancement, speech quality and intelligibility assessment, speech-signal processing, and deep learning.



TUAN VU HO received the B.E. degree from Vietnam National University, Ho Chi Minh, in 2015, the M.S. and Ph.D. degrees from the Japan Advanced Institute of Science and Technology (JAIST), in 2018 and 2021, respectively. His research interests include voice conversion, speech enhancement, speech-signal processing, and deep learning. He is currently working as a full-time researcher at Hitachi, Ltd.



MASATO AKAGI (Life Member, IEEE) received the B.E. degree from the Nagoya Institute of Technology, in 1979, and the M.E. and Ph.D. (Eng.) degrees from the Tokyo Institute of Technology, in 1981 and 1984, respectively. He joined the Electrical Communication Laboratories, Nippon Telegraph and Telephone Corporation (NTT), in 1984. From 1986 to 1990, he was with the ATR Auditory and Visual Perception Research Laboratories. Since 1992, he has been a Faculty Member with

the School of Information Science, JAIST. He is currently a professor emeritus. His research interests include speech perception, modeling of speech perception mechanisms in humans, and the signal processing of speech. He was a recipient of the IEICE Excellent Paper Award from IEICE, in 1987, the Best Paper Award from the Research Institute of Signal Processing, in 2009, and the Sato PRize for OUtstanding Papers from the Acoustical Society of Japan, in 1998, 2005, 2010, and 2011.



MASASHI UNOKI (Member, IEEE) received the M.S. and Ph.D. degrees in information science from the Japan Advanced Institute of Science and Technology (JAIST), in 1996 and 1999, respectively. His research interests include auditory-motivated signal processing and the modeling of auditory systems. He was a Japan Society for the Promotion of Science (JSPS) Research Fellow, from 1998 to 2001. He was associated with the ATR Human Information Processing Laboratories

as a Visiting Researcher, from 1999 to 2000. He was a Visiting Research Associate with the Centre for the Neural Basis of Hearing (CNBH), Department of Physiology, University of Cambridge, from 2000 to 2001. He has been a Faculty Member with the School of Information Science, JAIST, since 2001. He is currently a professor. He is a member of the Research Institute of Signal Processing (RISP), the Institute of Electronics, Information and Communication Engineers (IEICE) of Japan, and the Acoustical Society of America (ASA). He is also a member of the Acoustical Society of Japan (ASJ) and the International Speech Communication Association (ISCA). He received the Sato Prize from the ASJ, in 1999, 2010, and 2013 for an Outstanding Paper and the Yamashita Taro "Young Researcher" Prize from the Yamashita Taro Research Foundation, in 2005.

...