NOISY SPEECH ENHANCEMENT BASED ON PRIOR KNOWLEDGE ABOUT SPECTRAL ENVELOPE AND HARMONIC STRUCTURE

Takuya Yoshioka, Tomohiro Nakatani

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

Hiroshi G. Okuno

Graduate School of Informatics
Kyoto University
Yoshida-hommachi, Sakyo-ku
Kyoto 606-8501, Japan

ABSTRACT

This paper considers the enhancement of noisy speech. Earlier studies have revealed that an approach that enhances spectral envelopes by using prior knowledge about the all-pole (AP) model parameters of clean speech learnt from speech corpora is advantageous in terms of the amount of musical noise and speech distortion. This paper proposes a new speech enhancement method, in which harmonic structure enhancement is incorporated in learning-based spectral envelope enhancement to further improve performance. The harmonic structure is represented by using a harmonic Gaussian mixture model (GMM), which is parameterized by a voicing indicator and a fundamental frequency. The parameters of the AP model and the harmonic GMM are jointly estimated by maximum a posteriori estimation, thus enabling the enhancement of spectral envelopes and harmonic structures in a unified framework. The proposed method outperforms the spectral envelope enhancement approach by 0.85 dB in cepstral distance.

Index Terms— Speech enhancement, spectral envelope, harmonic structure, learning

1. INTRODUCTION

The quality and intelligibility of speech encountered in practical scenarios are often degraded by ambient noise. Thus, technologies for enhancing noisy speech are useful for a wide range of applications.

The research on noisy speech enhancement has been conducted over decades. The traditional spectral subtraction approach is likely to produce a lot of musical noise. In terms of amount of musical noise, the decision-directed approach [1] is advantageous over spectral subtraction. However, signals enhanced with this approach tend to sound reverberant [2].

Since the 1990s, a learning-based approach has been investigated extensively [3, 4]. This approach aims at enhancing the spectral envelopes of noisy speech by using prior knowledge about the spectral envelopes of clean speech acquired from speech corpora. The underlying assumption is that the distribution of clean speech spectral envelopes can be approximated by a finite number of prototypes since a spectral envelope of speech mainly encodes phonemic information. This approach seems to produce less musical noise than spectral subtraction. Moreover, the enhanced signals are unlikely to be greatly distorted.

On the other hand, some researchers have studied an approach that exploits the harmonic structure of speech [5]. Because the energy of voiced speech is concentrated at harmonic frequencies, local signal-to-noise ratios (SNRs) vary among time frames and frequency bins. Thus, at least in principle, the harmonicity-based approach is able to finely modify individual spectral components of an input signal depending on the local SNRs, resulting in the realization of a perceptually clear speech signal.

The above two approaches focus on different aspects of power spectra. The aim of the learning-based approach is to enhance spectral envelopes while the harmonicity-based approach capitalizes on the specific shape of the fine structures of speech spectra, i.e., the harmonic structure.

In this paper, we propose a speech enhancement method that combines the above two approaches. The proposed method enhances both spectral envelopes and harmonic structures by using both prior knowledge about clean speech spectral envelopes derived from speech corpora and a harmonic structure model. The enhancement of spectral envelopes and harmonic structures is performed in a unified framework. To realize this, a clean speech power spectrum is represented based on a source-filter model in which both the spectral envelope and harmonic structure are taken into account. The proposed method estimates all the parameters of the source-filter model from a noisy signal by means of maximum a posteriori (MAP)-estimation-based spectral matching, whereafter a Wiener filter derived from the estimated parameters is applied to the noisy signal to produce an enhanced signal.

Our source-filter model is designed as follows. The power spectrum of the articulatory filter is modeled by an all-pole (AP) spectrum, which is parameterized by linear prediction coefficients (LPCs) and a prediction residual power. A continuous prior distribution of the AP parameters (i.e., LPCs and prediction residual powers) acquired from speech corpora is used as a prior knowledge about clean speech spectral envelopes. On the other hand, the power spectrum of the excitation source is expressed by using a finite-sized dictionary of possible excitation spectra, each of which is modeled with the harmonic Gaussian mixture model (GMM) described in [6]. The spectral shapes of individual excitations correspond to unvoiced speech or different fundamental frequencies of voiced speech.

The use of a continuous prior distribution of AP parameters is the key to the success of the proposed method. Most existing AP-model-based spectral envelope enhancement methods use a codebook of AP parameters which means that the AP parameter domain is limited to a finite set. One may simply combine this AP parameter codebook and the above excitation dictionary to perform the joint spectral envelope and harmonic structure enhancement. However, this naive approach sometimes makes the use of the harmonic structure model futile or even harmful. This is because spectral envelopes and gains represented by the AP model are allowed to have one of a finite number of spectra and hence may become dominant during

spectrum matching. Using a continuous prior distribution rather than an AP parameter codebook is helpful in avoiding such situations.

It is important to note the necessity of prior knowledge about clean speech spectral shapes. It may be possible to enhance spectral envelopes and harmonic structures without such prior knowledge as in [7]. However, in our experience, the lack of a restriction on the AP parameters sometimes results in an incorrect fit of the AP spectra with the noise spectra or harmonic peaks.

The remainder of this paper is organized as follows. Sect. 2 reviews existing learning-based methods for spectral envelope enhancement. Sect. 3 describes the proposed method for joint enhancement of spectral envelopes and harmonic structures. Sect. 4 reports our experimental results, and our conclusion is provided in Sect. 5.

2. REVIEW OF ENVELOPE ENHANCEMENT METHODS

First, we define the goal of speech enhancement. Let $s_{n,l}$, $v_{n,l}$, and $x_{n,l}$ be the short time Fourier transform (STFT) coefficients of clean speech, noise, and noisy speech signals, respectively, at time frame n and frequency bin l. The noisy speech is the sum of the clean speech and noise, i.e, $x_{n,l} = s_{n,l} + v_{n,l}$. Now, let us use $\mathcal{S}_n$ to denote the set of all the clean speech STFT coefficients, or the complex spectrum, at time frame n as $\mathcal{S}_n = \{s_{n,0}, \cdots, s_{n,L-1}\}$. Noise complex spectrum $\mathcal{V}_n$ and noisy speech complex spectrum $\mathcal{X}_n$ are defined in the same fashion. Speech enhancement is a process of estimating clean speech $\mathcal{S}_n$ based solely on noisy speech $\mathcal{X}_0, \cdots, \mathcal{X}_n$ up to time frame n.

Spectral envelope enhancement methods can be classified according to envelope representations. The mel filter bank (MFB) and the AP model are typical methods for representing spectral envelopes [4]. An AP spectrum of order P is given by

$$\lambda^{\mathrm{AP}}(l;\theta) = \frac{e}{\left|1 - \sum_{k=1}^{P} a_k e^{-j\frac{2\pi l}{L}k}\right|^2}, \tag{1}$$

where $a_1, \cdots, a_P$ are LPCs, e is a prediction residual power, and $\theta = \{a_1, \cdots, a_P, e\}$. In this paper, we use the AP model because this model directly describes a power spectrum itself and hence can be easily combined with our harmonic structure model, which is defined on a linear frequency axis. Below, we review the AP-model-based approach that uses prior knowledge about the AP parameters of clean speech.

(A) In most existing methods, clean speech at each time frame is assumed to belong to one of the M_{E} states. Each state i is associated with a unique AP parameter set θ^i, thereby representing the power spectrum $\lambda^{\mathrm{AP}}(l;\theta^i)$, which is specific to this state. $\theta^1, \cdots, \theta^{M_{\mathrm{E}}}$ constitute a codebook of AP parameters. The ith state, or the ith codeword, has a prior probability of $P(i) = \pi^i$, where $\sum_{i=1}^{M_{\mathrm{E}}} \pi^i = 1$. At each time frame n, state i_n is activated according to this prior probability as

$$i_n \sim P(i). \tag{2}$$

The clean speech signal at time frame n is then generated by a Gaussian process whose power spectrum is that of the i_nth state, $\lambda^{\mathrm{AP}}(l;\theta^{i_n})$, as

$$s_{n,l}|i_n \sim \mathcal{N}_{\mathbb{C}}(s;0,\lambda^{\mathrm{AP}}(l;\theta^{i_n})) \tag{3}$$

The prior probability π^i and the AP parameter set θ^i for each i are learnt from speech corpora in advance. This is accomplished by using the LPC-VQ and the EM algorithm.

Then, enhancement is performed by using the clean speech model given by (2) and (3). Let $\hat{s}_{n,l}$ denote an enhanced spectrum. $\hat{s}_{n,l}$ is obtained as the minimum mean square error (MMSE) estimate of clean speech $s_{n,l}$, which is calculated as

$$\hat{s}_{n,l} = \sum_{i_n=1}^{M_{\mathrm{E}}} p(i_n|\mathcal{X}_n) G(\lambda^{\mathrm{AP}}(l;\theta^{i_n}), \lambda_n^{\mathrm{V}}(l)) x_{n,l}, \tag{4}$$

where $\lambda_n^{\mathrm{V}}(l)$ is the noise power spectrum at time frame n. $p(i_n|\mathcal{X}_n)$ is the posterior probability of state i_n, and $G(\lambda^{\mathrm{AP}}(l;\theta^{i_n}), \lambda_n^{\mathrm{V}}(l))$ is the Wiener filter associated with state i_n, which is defined as

$$G(\lambda^{\mathrm{AP}}(l;\theta^{i_n}), \lambda_n^{\mathrm{V}}(l)) = \frac{\lambda^{\mathrm{AP}}(l;\theta^{i_n})}{\lambda^{\mathrm{AP}}(l;\theta^{i_n}) + \lambda_n^{\mathrm{V}}(l)}. \tag{5}$$

(4) means that the enhanced speech $\hat{s}_{n,l}$ is the weighted sum of the state-conditional Wiener filter outputs. Although it may be possible to adaptively estimate the noise power spectrum $\lambda_n^{\mathrm{V}}(l)$ by modeling noise with a GMM or the like, in this paper, $\lambda_n^{\mathrm{V}}(l)$ is assumed to be estimated in advance during non-speech periods.

3. PROPOSED METHOD FOR JOINT ENVELOPE AND HARMONIC STRUCTURE ENHANCEMENT

Now, we describe the proposed speech enhancement method. In Sect. 2, a clean speech power spectrum is modeled by an AP spectrum. From the viewpoint of the source-filter model, this is equivalent to assuming the excitation source to be white noise. However, a voiced speech power spectrum has a harmonic structure. To take this harmonic structure into account, we model a clean speech power spectrum, denoted by $\lambda^{\mathrm{S}}(l;\phi)$, as the product of an all-pole spectrum $\lambda^{\mathrm{AP}}(l;\theta)$ and a harmonic spectrum $\lambda^{\mathrm{H}}(l;\eta,\mu)$:

$$\lambda^{\mathrm{S}}(l;\phi) = \lambda^{\mathrm{AP}}(l;\theta)\lambda^{\mathrm{H}}(l;\eta,\mu), \tag{6}$$

where $\phi = \{\theta, \mu, \eta\}$. Parameter η is a voicing indicator, which indicates the voiced and unvoiced states at $\eta = 1$ and $\eta = 0$, respectively. Parameter μ is a fundamental frequency of voiced speech, which is disabled when $\eta = 0$.

The harmonic spectrum $\lambda^{\mathrm{H}}(l;\eta,\mu)$ is defined by using the harmonic GMM proposed in [6]. To be more precise, $\lambda^{\mathrm{H}}(l;\eta,\mu)$ is define as

$$\lambda^{\mathrm{H}}(l;\eta,\mu) = \begin{cases} \sum_{k=\infty}^{\infty} \exp\Big(-\frac{(2\pi l/L - k\mu)^2}{\xi}\Big) & \text{if } \eta = 1 \\ 1 & \text{if } \eta = 0 \end{cases}, \tag{7}$$

where ξ is a prescribed constant representing the spectral spread of each harmonic component. Fig. 1 shows an example of a voiced speech power spectrum given by (6). As seen in the figure, a certain spectral floor is introduced to $\lambda^{\mathrm{H}}(l;\eta,\mu)$.

As we have seen, with the spectral envelope enhancement approach, a clean speech signal is assumed to be generated according to (A). With the proposed method, we make the following assumptions about the clean speech generation process.

(B) The articulatory filter of speech is assumed to belong to one of the M_{E} states at each time frame. Unlike in (A), each state i is associated with a unique prior distribution, $p(\theta|i)$, of AP parameter set θ. Hence, state i represents a distribution of spectral envelopes rather than a power spectrum prototype. At each time frame n, articulatory state i_n is activated according to prior probability $P(i) = \pi^i$ as in

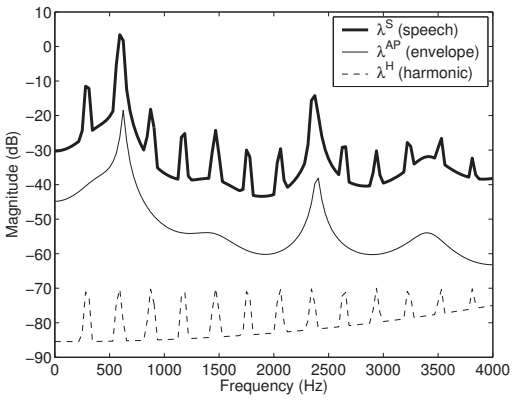

Fig. 1. An example of the speech power spectrum $\lambda^{\mathrm{S}}(l;\phi)$ along with the envelope part $\lambda^{\mathrm{AP}}(l;\theta)$ and the harmonic part $\lambda^{\mathrm{H}}(l;\eta,\mu)$. Individual plots are separated to improve visibility.

(2). Then, an AP parameter set, θ_n, for time frame n is drawn from the state-conditional prior distribution $p(\theta|i_n)$:

$$\theta_n \sim p(\theta|i_n). \tag{8}$$

Note that θ_n is now allowed to have any values unlike in (A), where the domain of θ_n is limited to a finite set.

On the other hand, the excitation source is also assumed to take one of the M_{H} states at each time frame. Each state j is associated with a unique voicing indicator η^j and a fundamental frequency μ^j, and therefore represents the harmonic spectrum, $\lambda^{\mathrm{H}}(l;\eta^j,\mu^j)$, specific to this state. The M_{H}th state corresponds to the unvoiced state, i.e., $\eta^{M_{\mathrm{H}}}=0$. $\eta^j=1$ for state $j=1,\cdots,M_{\mathrm{H}}-1$, and these states correspond to different fundamental frequencies $\mu^1,\cdots,\mu^{M_{\mathrm{H}}-1}$. State j is also associated with a prior probability of $P(j)=\omega^j$, where $\sum_{j=1}^{M_{\mathrm{H}}}\omega^j=1$. At each time frame n, excitation state j_n is activated according to this prior probability as

$$j_n \sim P(j). \tag{9}$$

Finally, the clean speech signal at time frame n is generated from a complex Gaussian distribution whose power spectrum is given by $\lambda^{\mathrm{S}}(l;\phi_n=\{\theta_n,\eta^{j_n},\mu^{j_n}\})=\lambda^{\mathrm{AP}}(l;\theta_n)\lambda^{\mathrm{H}}(l;\eta^{j_n},\mu^{j_n})$:

$$s_{n,l}|\theta_n,j_n \sim \mathcal{N}_{\mathbb{C}}(0,\lambda^{\mathrm{S}}(l;\phi_n)). \tag{10}$$

The above is the assumed clean speech generation process. □

In Subsect. 3.1, we define the state-conditional prior distribution, $p(\theta|i)$, of an AP parameter set. In Subsect. 3.2, we present an overview of the proposed speech enhancement algorithm, which is detailed in Subsect. 3.3.

3.1. Design of prior distributions of AP parameter set

We begin by formalizing state-conditional prior distribution $p(\theta|i)$. Let us vectorize LPC coefficients as $\boldsymbol{a}=[a_1,\cdots,a_P]^T$. We define $p(\theta|i)$ by using a normal-inverse gamma distribution as

$$p(\theta|i) = \mathcal{N}(\boldsymbol{a};\boldsymbol{\nu}^i,(\zeta^i\Xi^i/e)^{-1})\mathcal{IG}(e;\rho^i/2,\gamma^i/2), \tag{11}$$

where $\mathcal{IG}$ denotes an inverse gamma distribution. The use of the normal-inverse gamma distribution leads to a simple algorithm for MAP estimation of the AP parameters. In (11), ζ^i and ρ^i are scalars indicating the degrees of reliability of LPC vector $\boldsymbol{a}$ and prediction residual power e, respectively; Ξ^i is a normalized precision matrix of $\boldsymbol{a}$; γ^i is a scalar determining the magnitude of e.

The values of these hyperparameters may be estimated directly from speech corpora. Instead, we set the hyperparameter values by using an existing codebook of AP parameter sets, $\Theta=\{\theta^1,\cdots,\theta^{M_{\mathrm{E}}}\}$, according to the following procedure, which is derived from the Bayesian estimation rule of the AP parameters. First, the degrees of reliability, ζ^i and ρ^i, are given by $\zeta^i=\rho^i=\alpha$, where α is a prescribed reliability factor. A larger α value indicates a tighter prior distribution; at the extreme of $\alpha=\infty$, the prior distribution reduces to the codebook Θ. The LPC vector mean $\boldsymbol{\nu}^i$ is set at the LPC vector codeword $\boldsymbol{a}^i$. Ξ^i is given as a Toeplitz matrix consisting of the P auto-correlation coefficients that the AP parameter codeword θ^i represents. Finally, γ^i is given by $\gamma^i=e^i\rho^i$.

3.2. Overview of proposed algorithm

The proposed speech enhancement algorithm can be described as follows.

(s1) For each pair of articulatory and excitation states $(i_n,j_n)=(1,1),\cdots,(M_{\mathrm{E}},M_{\mathrm{H}})$, the MAP estimate of AP parameter set θ_n is calculated as

$$\hat{\theta}_n^{i_n,j_n} = \operatorname*{argmax}_{\theta_n} p(\theta_n|\mathcal{X}_n,i_n,j_n). \tag{12}$$

This step is described in detail in the next subsection.

(s2) For each state pair (i_n,j_n), the posterior probability of the state pair is calculated. Since the exact posterior probability is difficult to obtain, in practice, it is approximated by $p(i_n,j_n,\hat{\theta}_n^{i_n,j_n}|\mathcal{X}_n)$, which can be easily calculated.

(s3) For each state pair (i_n,j_n), a state-pair-conditional Wiener filter $G(\lambda^{\mathrm{S}}(l;\hat{\phi}_n^{i_n,j_n}),\lambda_n^{\mathrm{V}}(l))$ is calculated, where $\hat{\phi}_n^{i_n,j_n}=\{\hat{\theta}_n^{i_n,j_n},\eta^{j_n},\mu^{j_n}\}$ and function G is given by (5). Then, enhanced speech $\hat{s}_{n,l}$ is obtained as the weighted sum of the state-pair-conditional Wiener filter outputs as

$$\hat{s}_{n,l} = \sum_{i_n=1}^{M_{\mathrm{E}}}\sum_{j_n=1}^{M_{\mathrm{H}}} P(i_n,j_n,\hat{\theta}_n^{i_n,j_n}|\mathcal{X}_n)\hat{s}_{n,l}^{i_n,j_n}, \tag{13}$$

$$\text{where}\quad \hat{s}_{n,l}^{i_n,j_n} = G(\lambda^{\mathrm{S}}(l;\hat{\phi}_n^{i_n,j_n}),\lambda_n^{\mathrm{V}}(l))x_{n,l}. \tag{14}$$

In our implementation, a pruning process is incorporated in the above algorithm to reduce computational complexity.

3.3. MAP estimation of AP parameters

In this subsection, we describe an algorithm for step (s1) that calculates the state-pair-conditional MAP estimate, $\hat{\theta}_n^{i_n,j_n}$, of an AP parameter set. An EM algorithm is used to realize the MAP estimation, where the latent variables are clean speech STFT coefficients $s_{n,0},\cdots,s_{n,L-1}$. In this paper, we omit the derivation and merely show the formulae for the E- and M-steps.

E-step The state-pair-conditional estimate of clean speech, $\hat{s}_{n,l}^{i_n,j_n}$, and the associated expected square error, denoted by $\epsilon_{n,l}^{i_n,j_n}$, are calculated by using tentative MAP estimate $\hat{\theta}_n^{i_n,j_n}$. $\hat{s}_{n,l}^{i_n,j_n}$ is given by (14) while $\epsilon_{n,l}^{i_n,j_n}$ is defined as

$$\epsilon_{n,l}^{i_n,j_n} = G(\lambda^{\mathrm{S}}(l;\hat{\phi}_n^{i_n,j_n}),\lambda_n^{\mathrm{V}}(l))\lambda_n^{\mathrm{V}}(l). \tag{15}$$

Table 1. Summary of cepstral distances (dB). For each input SNR, the smallest CDs are indicated by boldface type.

Input SNR	15 dB	10 dB	5 dB	0 dB	Avg.
Noisy	6.31	7.06	7.67	8.13	7.29
Proposed, $\alpha = 0.5$	**3.30**	**3.65**	**4.19**	4.99	**4.03**
Proposed, $\alpha = 1.0$	3.52	3.81	4.27	**4.96**	4.14
Proposed, $\alpha = 4.0$	4.54	4.77	5.14	5.71	5.04
Envelope, $M_{\mathrm{E}} = 64$	4.28	4.57	5.00	5.67	4.88
Envelope, $M_{\mathrm{E}} = 512$	4.18	4.43	4.84	5.56	4.75

M-step A tentative estimate of the clean speech spectral envelope is calculated as $\hat{u}_{n,l}^{i_n,j_n} = (|\hat{s}_{n,l}^{i_n,j_n}|^2 + \epsilon_{n,l}^{i_n,j_n})/\lambda^{\mathrm{H}}(l;\eta^{j_n},\mu^{j_n})$. Then, $\hat{u}_{n,l}^{i_n,j_n}$ is converted to the corresponding auto-correlation coefficients via an inverse fast Fourier transform (IFFT):

$$\hat{r}_n^{i_n,j_n}[0],\cdots,\hat{r}_n^{i_n,j_n}[L-1] = \mathrm{IFFT}_L(\hat{u}_{n,l}^{i_n,j_n}), \quad (16)$$

where $\mathrm{IFFT}_L(\cdot)$ denotes an L-point IFFT. Subsequently, an auto-correlation matrix $R_n^{i_n,j_n}$ and an auto-correlation vector $\boldsymbol{r}_n^{i_n,j_n}$ modified by the prior distribution for the i_nth articulatory state are calculated as

$$R_n^{i_n,j_n} = \frac{L\hat{R}_n^{i_n,j_n} + \zeta^{i_n}\Xi^{i_n}}{L+\zeta^{i_n}} \quad (17)$$

$$\boldsymbol{r}_n^{i_n,j_n} = \frac{L\hat{\boldsymbol{r}}_n^{i_n,j_n} + \zeta^{i_n}\Xi^{i_n}\boldsymbol{\nu}^{i_n}}{L+\zeta^{i_n}}, \quad (18)$$

where $\hat{R}_n^{i_n,j_n}$ is the symmetric Toeplitz matrix consisting of $\hat{r}_n^{i_n,j_n}[0],\cdots,\hat{r}_n^{i_n,j_n}[P-1]$ and $\hat{\boldsymbol{r}}_n^{i_n,j_n} = [\hat{r}_n^{i_n,j_n}[1],\cdots,\hat{r}_n^{i_n,j_n}[P]]^T$. The tentative MAP estimate of the LPC vector and prediction residual power, $\hat{\boldsymbol{a}}_n^{i_n,j_n}$ and $\hat{e}_n^{i_n,j_n}$, are updated by solving the following Yule-Walker equation as follows:

$$\begin{bmatrix} r_n^{i_n,j_n}[0] & (\boldsymbol{r}_n^{i_n,j_n})^T \\ \boldsymbol{r}_n^{i_n,j_n} & R_n^{i_n,j_n} \end{bmatrix} \begin{bmatrix} 1 \\ -\hat{\boldsymbol{a}}_n^{i_n,j_n} \end{bmatrix} = \begin{bmatrix} \hat{e}_n^{i_n,j_n} \\ 0 \end{bmatrix}, \quad (19)$$

where $r_n^{i_n,j_n}[0]$ is the diagonal component of the Toeplitz matrix $R_n^{i_n,j_n}$, which may be considered as the 0th-lag modified auto-correlation coefficient.

4. EXPERIMENTAL RESULTS

We conducted experiments to evaluate the effectiveness of the proposed method. The Aurora-2 clean training set was used for learning the prior distribution of the AP parameters of clean speech. For our evaluation, we used the Aurora-2 test set A, which consists of speech data contaminated by four types of noise (i.e., subway, babble, car, and exhibition) at various SNRs.

The proposed method was implemented with the design described below. The STFT was performed with a 200-point hamming window, an 80-point shift, and a 256-point FFT. The AP model order was set at 10. The number of articulatory states, M_{E}, was 64. The excitation spectrum dictionary contained $M_{\mathrm{H}} = 52$ spectra. The fundamental frequencies of these spectra were between 90 and 350 Hz, where adjacent fundamental frequencies were separated by 50 cent. Several values were tested for the reliability factor, α, of the AP-parameter prior distributions.

Table 1 lists the average cepstral distances of noisy and enhanced speech from clean speech for each input SNR. Also shown are results obtained with the spectral envelope enhancement method, which uses an AP parameter codebook and is described in Sect. 2. It can be seen that the proposed method, which enhances both spectral envelopes and harmonic structures, achieved better speech enhancement than the spectral envelope enhancement method. In addition, we can make the following two observations.

- Letting the reliability factor α be very large led to performance degradation. As explained earlier, a larger α value indicates a tighter prior distribution of an AP parameter set. Therefore, this tendency confirms the necessity of using a continuous prior distribution rather than an AP parameter codebook when combining spectral envelope enhancement and harmonic structure enhancement.
- The proposed method also outperformed the spectral envelope enhancement method with a larger-sized codebook ($M_{\mathrm{E}} = 512$). This indicates the essential advantage of exploiting both the harmonic structure and prior knowledge about spectral envelopes in noisy speech enhancement.

5. CONCLUSION

We have presented a new speech enhancement method that enhances both spectral envelopes and harmonic structures by using a prior distribution of clean speech AP parameters and a dictionary of harmonic spectra. Our experimental results revealed the advantage of exploiting both the harmonic structure and prior knowledge about the clean speech spectral envelopes. The results also indicated the importance of using a continuous prior distribution rather than an AP parameter codebook.

Harmonic structures will also be useful for noise spectrum adaptation. Thus, future work will include the incorporation of noise spectrum adaptation into the proposed method.

6. REFERENCES

[1] Y. Ephraim, "Speech enhancement using a minimum mean square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 32, no. 6, pp. 1109–1121, 1984.

[2] C. Plapous, C. Marro, and P. Scalart, "Improved signal-to-noise ratio estimation for speech enhancement," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 6, pp. 2098–2108, 2006.

[3] Y. Ephraim, "Statistical-model-based speech enhancement systems," *Proc. IEEE*, vol. 80, no. 10, pp. 1526–1555, 1992.

[4] J. C. Segura, A. de la Torre, M. C. Benitez, and A. M. Peinado, "Model-based compensation of the additive noise for continuous speech recognition. Experiments using the AURORA II databased and tasks," in *Proc. Eurospeech*, 2001, pp. 221–224.

[5] H.-G Kim, M. Schwab, N. Moreau, and T. Sikora, "Speech enhancement of noisy speech using log-spectral amplitude estimator and harmonic tunneling," in *Int'l Worksh. Acoust. Echo, Noise Contr.*, 2003, pp. 119–122.

[6] H. Kameoka, T. Nishimoto, and S. Sagayama, "Multi-pitch trajectory estimation of concurrent speech based on harmonic GMM and nonlinear Kalman filtering," in *Proc. Interspeech*, 2004, pp. 2433–2466.

[7] Q. Yan, S. Vaseghi, E. Zavarehei, B. Milner, J. Darch, P. White, and I. Andrianakis, "Kalman tracking of linear predictor and harmonic noise models for noisy speech enhancement," *Comp. Speech, Lang.*, vol. 22, no. 1, pp. 69–83, 2008.