DNN-based Statistical Parametric Speech Synthesis Incorporating Non-negative Matrix Factorization

Shunsuke Goto, Daisuke Saito, Nobuaki Minematsu
Graduate School of Engineering, The University of Tokyo, Japan
E-mail: {goto, dsk_saito, mine}@gavo.t.u-tokyo.ac.jp

***Abstract*—This paper proposes a novel approach of DNN-based statistical parametric speech synthesis where non-negative matrix factorization (NMF) is effectively utilized. In statistical parametric speech synthesis, Mel-frequency cepstrum is often employed for acoustic features. However, it represents a spectral envelope as a linear combination of fixed envelope curves (sines and cosines), and the envelope predicted by a DNN-based acoustic model loses its fine structure. On the other hand, in NMF, multiple spectral envelopes (spectrogram) are decomposed into two factors; spectral bases and their activity patterns (activation). Since the obtained bases keep the fine structure of envelopes, the remaining factor, i.e. activation can be employed for acoustic features. Due to its sparseness, the spectral envelope obtained by the predicted activation also keeps fine structure. In this study, activation derived from NMF is utilized for spectral representation, and DNN-based text-to-speech synthesis incorporating NMF is proposed. In addition, this framework can potentially incorporate some applications of NMF, such as bandwidth expansion, voice conversion, or noise reduction. In this study, bandwidth expansion is achieved, and experimental results demonstrate that the proposed method can generate more natural spectral parameters especially in 48 kHz sampling rate, and that 16 kHz-to-48 kHz bandwidth expansion, where natural synthetic speech is produced, is achieved in the proposed framework.**

I. Introduction

Text-to-speech (TTS) synthesis is a technology that converts texts into the corresponding speech waveforms. One of the conventional approach of TTS is statistical parametric speech synthesis (SPSS) [1], [2]. SPSS is based on source-filter model, which models speech as a combination of a sound source (vocal cords), and a filter (vocal tract). Fundamental frequency (F_0) is the acoustic feature for a sound source, and Mel-frequency cepstrum (Mel-cepstrum) is that for a filter. These acoustic features are predicted from linguistic features by hidden Markov models (HMMs) or deep neural networks (DNNs), and speech waveforms are generated using vocoder. However, recently, thanks to the advancement of DNNs, it has been possible to directly predict amplitude spectrogram, or speech waveforms from linguistic features or texts [3], [4], [5], [6]. Unlike SPSS, these methods are not based on the speech generation model such as source-filter model. Though some of them can generate very natural speech waveforms, they need a large amount of speech data, and it is usually difficult to train these models. On the other hand, in SPSS, which is based on source-filter model, the generated speech sounds not so natural, but it has some advantages that it requires fewer data, and the control of speech is relatively easy. Therefore, it is still important to improve the quality of generated speech in SPSS.

There has been previous work trying to improve the performance of SPSS. In [7], and [8], to capture the time dependency, recurrent neural networks (RNNs), or long short-term memory (LSTM) is employed for acoustic models representing the relationship between linguistic and acoustic features. In [9] and [10], more precise excitation model is proposed. However, Mel-cepstrum is still employed for spectral envelope modeling. Although there are some attempts trying to model spectral envelope directly [11], [12], the alternative acoustic feature needs to be considered.

Mel-cepstrum, which is the inverse Fourier transform of the logarithmic spectrum on a Mel-scaled frequency, is widely used as acoustic features to represent spectral envelopes. It can efficiently model the spectral envelopes by a small number of coefficients. However, it represents a spectral envelope as a linear combination of fixed envelope curves (sines and cosines), and the envelope predicted by a DNN-based acoustic model loses its fine structure. On the other hand, there are some attempts to model spectral envelopes without intermediate representation such as Mel-cepstrum [11], [12]. In these methods, while the fine structure of spectral envelopes can be kept, the modeling is not easy because the number of the coefficients is large.

Acoustic features to model spectral envelopes should satisfy two requirements; to keep the fine structure of envelopes and to represent the envelopes by a small number of parameters. To address these problem, we focus on non-negative matrix factorization (NMF) [13]. NMF is a set of algorithms to factorize a non-negative matrix into the multiplication of two non-negative matrices. By applying NMF to multiple spectral envelopes (spectrogram), we obtain two factors; spectral bases and their activity patterns (activation). Each of the envelopes is represented as a weighted linear combination of the spectral bases. With the help of non-negativity and sparseness, NMF can make the obtained spectral bases keep their fine structure. Because of the sparseness of the activation, the spectral envelope obtained by the predicted activation also keeps fine structure. Moreover, the dimension of activity patterns is lower than that of spectral envelopes. Therefore, NMF can balance the two requirements, and it is suitable for extraction of acoustic features.

In this paper, we propose a combination of NMF with DNN-based speech synthesis. In the proposed method, activity

patterns derived from NMF are employed for acoustic features which should be modeled by DNNs. To make DNNs model the sparse activity patterns properly, we employ a suitable loss function for them, which can be derived from the definition of the generalized Kullback-Leibler divergence (KLD).

NMF is widely used for voice conversion [14], noise reduction [15], etc. by operating the constructed spectral bases while keeping the activity patterns. Since the proposed method models the relationship between linguistic features and the activity patterns, these applications and TTS can be combined flexibly. In this paper, the proposed framework is combined with bandwidth expansion [16]. By preparing a small amount of wide-band samples, the proposed method can generate the wide-band synthetic speech even when the acoustic model is constructed by narrow-band samples.

The rest of the paper is organized as follows: Section II explains the modeling of spectral parameters in TTS. Section III describes the basic idea of NMF and explains the proposed DNN-based TTS incorporating NMF. Section IV presents experimental evaluations. Finally, Section V concludes the paper.

II. Modeling of Spectral Parameters in TTS

A. Mel-cepstrum as Intermediate Features

In DNN-based SPSS, DNNs are employed for acoustic models which represent the relationship between linguistic and acoustic features [2]. Linguistic features, which consist of binary or numerical answers to linguistic questions, are obtained from text analysis. Acoustic features should capture both the vocal tract and vocal cords information, which respectively correspond to spectral envelopes and fundamental frequencies. Hence, the features consist of Mel-cepstrum, F_0, and band aperiodicity measures. Mel-cepstrum coefficients are the intermediate features for representing the spectral envelope.

In general, to train a DNN-based acoustic model, the mean squared error (MSE) is adopted as the objective criterion which should be minimized. Let T and D be the number of frames and the dimension of the features, respectively. MSE is defined as

$$\mathcal{L}_{\mathrm{MSE}}\left(\boldsymbol{y} \mid \hat{\boldsymbol{y}}\right) = \frac{1}{2}\sum_{t=1}^{T}\sum_{d=1}^{D}\left(y_{t,d} - \hat{y}_{t,d}\right)^2, \tag{1}$$

where $\boldsymbol{y}$ is a sequence of acoustic features obtained from natural speech, and $\hat{\boldsymbol{y}}$ is that predicted by DNNs.

B. Spectal Modeling without Intermediate Features

In Mel-cepstrum, spectral representation in frequency domain is converted into cepstral representation. It causes the loss of the fine structure of spectral envelopes. To avoid this problem, in some attempts, spectral envelopes are modeled in frequency domain without intermediate representation such as Mel-cepstrum. However, the number of coefficients for spectral envelopes depends on the size of frame windows, and it is usually larger than 512. Hence, it is essential to treat the high dimensional features appropriately. In [11], restricted Boltzmann machines (RBMs) have been used to model the output probabilities of the high-dimensional features in HMM. RBMs can capture the strong correlations among the coefficients in the spectral envelopes rather than Gaussian mixture models. In [12], autoencoder is utilized to obtain the suitable representation for spectral envelopes, and the obtained features are predicted from the linguistic features by DNNs.

On the other hand, instead of decomposition into spectral envelopes and fundamental frequencies, the prediction of amplitude spectra is adopted [3]. Because the waveforms can be derived from the predicted amplitude spectra by phase recovery, a vocoder need not be used. The study shows that KLD is more suitable than MSE as the objective criterion for predicting the amplitude spectrogram. Let $\boldsymbol{y}$ and $\hat{\boldsymbol{y}}$ be the reference sequence and the predicted one, respectively. The KLD between them is defined as

$$\mathcal{L}_{\mathrm{KLD}}\left(\boldsymbol{y} \mid \hat{\boldsymbol{y}}\right) = \sum_{t=1}^{T}\sum_{d=1}^{D} y_{t,d}\log\frac{y_{t,d}}{\hat{y}_{t,d}} - y_{t,d} + \hat{y}_{t,d}. \tag{2}$$

III. DNN-based TTS Incorporating NMF

A. Non-negative Matrix Factorization

Non-negative matrix factorization (NMF) is a group of algorithms where matrix $\boldsymbol{Y} = (y_{k,n})_{K\times N}$ is factorized into two matrices $\boldsymbol{H} = (h_{k,m})_{K\times M}$, $\boldsymbol{U} = (u_{m,n})_{M\times N}$, with the property that all the matrices have no negative elements. That is,

$$\boldsymbol{Y} \simeq \boldsymbol{H}\boldsymbol{U}. \tag{3}$$

Usually, $K \gg M$, and $N \gg M$. The matrix $\boldsymbol{H}$ is called *exemplar* or *dictionary*, and $\boldsymbol{U}$ is called *activation*. In the modeling of spectra with NMF, a spectrum at the n-th frame is represented as a linear combination of basis spectra $\boldsymbol{h}_1, \ldots, \boldsymbol{h}_M$ as follows;

$$\boldsymbol{y}_n \simeq \sum_{m=1}^{M} \boldsymbol{h}_m u_{m,n} = \boldsymbol{H}\boldsymbol{u}_n. \tag{4}$$

Because of the non-negativity of NMF, activation tends to be sparse 1, and it is possible to make activation more sparse with with sparseness constraints [17]. In the proposed method, the activation $\boldsymbol{u}_n$ which is extracted from the spectral envelope $\boldsymbol{y}_n$ is the acoustic feature for spectral envelope modeling.

To find an approximate factorization $\boldsymbol{Y} \simeq \boldsymbol{H}\boldsymbol{U}$, the elements of two matrices $\boldsymbol{H}, \boldsymbol{U}$ are iteratively updated based on a cost function. In this paper, based on [14], KLD is employed as a cost funtion for the amplitude spectrogram. The divergence between x and y is

$$\mathcal{D}_{KL}(y \mid x) = y\,\log\frac{y}{x} - y + x. \tag{5}$$

To minimize KLD, the elements of the two matrices are

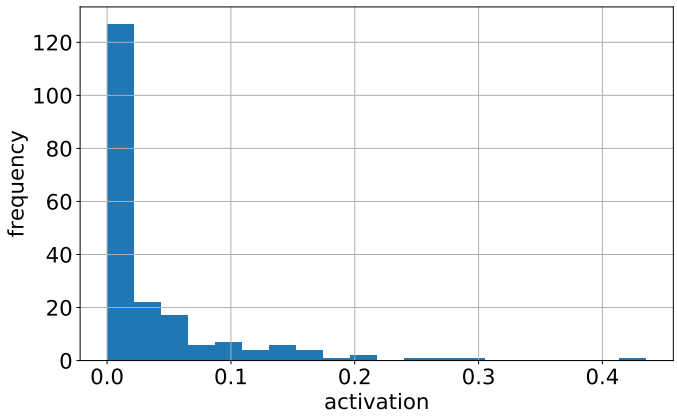

Fig. 1. Histogram of activation at one frame.

updated as follows [13];

$$h'_{k,m} \leftarrow h_{k,m} \frac{\sum_n y_{k,n} u_{m,n} / x_{k,n}}{\sum_n u_{m,n}}, \quad (6)$$

$$u'_{m,n} \leftarrow u_{m,n} \frac{\sum_n y_{k,n} h_{k,m} / x_{k,n}}{\sum_k h_{k,m}}, \quad (7)$$

$$x_{k,n} = \sum_m h_{k,m} u_{m,n}. \quad (8)$$

B. Activation as Intermediate Features

In Mel-cepstrum, a spectral envelope is represented as a linear combination of fixed envelope curves (sines and cosines). Similarly, in NMF, a spectrum is also represented as a linear combination of basis spectra. Therefore, in both of them, spectral envelopes are represented efficiently.

However, in NMF, spectral bases are obtained flexibly by the decomposition of the spectrogram and each basis spectrum has fine structure. Since the activation of NMF tends to be sparse, the spectral envelope which is obtained from the activation predicted by the acoustic model would also have fine strcuture.

In addition, the bases of NMF can be obtained depending on the speech data. For example, bases depeding on a speaker or a sampling rate can be made. Therefore, this representation will incorporate some applications such as voice conversion [14] or bandwidth expansion [16].

C. Prediction of Activation

The overview of the proposed TTS scheme incorporating NMF is shown in Figure 2. $\boldsymbol{Y}$, the amplitude spectrogram obtained by training speech data, is factorized into two matrices $\boldsymbol{H}$ and $\boldsymbol{U}$, and then a DNN-based acoustic model representing the relationship between the linguistic and acoustic features ($\boldsymbol{U}$) is trained. By multiplying the dictionaries $\boldsymbol{H}$ and the predicted activation $\boldsymbol{U'}$, the predicted amplitude spectrogram $\boldsymbol{Y'}$ is obtained.

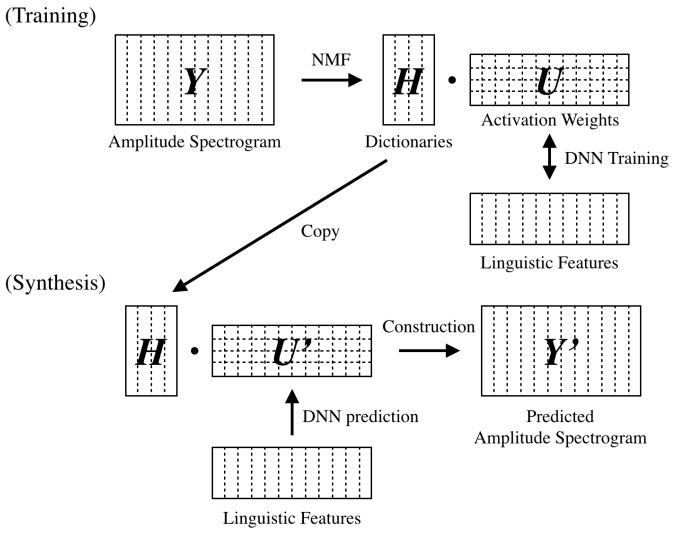

Fig. 2. Overview of statistical parametric speech synthesis incorporating NMF.

NMF has a property that it produces a sparse representation of data. Since it has a large number of elements whose value is zero, MSE is not a suitable loss function for modeling such a representation. Hence, we derive a suitable loss function for it.

Let $\boldsymbol{u'} = [u'_1, u'_2, \ldots, u'_M]$ be an activation at one frame, and $c = \sum_{m=1}^{M} u'_m$. By dividing each bin of $\boldsymbol{u'}$ by the sum c, the normalized activation, $\boldsymbol{u} = [u_1, u_2, \ldots, u_M]$ is obtained as

$$\boldsymbol{u} = \frac{\boldsymbol{u'}}{c}. \quad (9)$$

Since the sum of $\boldsymbol{u}$ is 1, $\boldsymbol{u}$ can be regarded as categorical distribution. The KLD between the predicted activation $\hat{\boldsymbol{u'}}$ and the actual activation $\boldsymbol{u'}$ is

$$\mathcal{D}_{KL}(\boldsymbol{u'} \mid \hat{\boldsymbol{u}}') = \sum_m \left(c u_m \log \frac{c u_m}{\hat{c} \hat{u}_m} - c u_m + \hat{c} \hat{u}_m \right)$$
$$= c \left\{ -\sum_m u_m \log \hat{u}_m + \left(\frac{\hat{c}}{c} - \log \frac{\hat{c}}{c} - 1 \right) + \sum_m u_m \log u_m \right\}$$
$$= c \left\{ \mathcal{D}_{CE}(\boldsymbol{u} \mid \hat{\boldsymbol{u}}) + \mathcal{D}_{IS}(\hat{c} \mid c) - \mathcal{D}_{CE}(\boldsymbol{u} \mid \boldsymbol{u}) \right\}, \quad (10)$$

where $\mathcal{D}_{CE}$ and $\mathcal{D}_{IS}$ denote the cross-entropy and the Itakura-Saito divergence, respectively.

Therefore, the minimization of the KLD is equal to that of the sum of the cross-entropy between $\hat{\boldsymbol{u}}$ and $\boldsymbol{u}$, and the Dual-Itakura-Saito divergence (D-ISD)[1] between $\hat{c}$ and c;

$$\mathcal{L}_{KL} = -\sum_m u_m \log \hat{u}_m + \left(\frac{\hat{c}}{c} - \log \frac{\hat{c}}{c} - 1 \right). \quad (11)$$

In the proposed method, the above $\mathcal{L}_{KL}$ is employed for the loss function.

[1]The estimator and the reference in Itakura-Saito divergence is switched with each other in this divergence.

TABLE I
EXPERIMENTAL CONDITIONS.

Model	Output	Dimension	Normalization	Loss Function	Output Layer
MCEP	Mel-cepstrum	180 (mcep+Δ+ΔΔ)	Mean: 0, Var: 1	MSE	linear
SP	spectrum	513 (16 kHz), 1025 (48 kHz)	Min: 0.01, Max: 0.99	KLD	sigmoid
LogSP	log spectrum	513 (16 kHz), 1025 (48 kHz)	Mean: 0, Var: 1	MSE	linear
ACT	activation	201 (norm-act: 200 + power: 1)	Sum:1 + None	CE + D-ISD	softmax + softplus

D. Bandwidth Expansion with the Proposed Method

As described in Section III-B, NMF is widely used for voice conversion [14], noise reduction [15], etc. by operating the constructed spectral bases while keeping the activity patterns. In the proposed method, by replacing the spectral bases from narrow-band samples with those from wide-band samples, bandwidth expansion is achieved. By using pairs of utterances from narrow-band and wide-band samples, parallel dictionaries are constructed on the assumption that the same activation patterns are obtained even from the different bandwidth samples. The wide-band speech samples are composed of the replaced dictionary and the activation patterns predicted by the DNN trained with narrow-band samples.

IV. EXPERIMENTS

A. Experimental Conditions

Evaluations were carried out using phonetically balanced 503 sentences from ATR Japanese speech database [18]. Speech samples uttered by a male speaker in the demo script of HTS were used[2]. We used 450 sentences for training, and 53 sentences for evaluation. Samples in two different sampling rates were prepared; 16 kHz and 48 kHz. The data in 16 kHz sampling rate were downsampled from their original samples in 48 kHz sampling rate. WORLD was utilized to obtain spectral envelopes, F_0, and band aperiodicity measures, and to generate a waveform from the predicted acoustic features [19].

We constructed three baseline models (MCEP, SP, LogSP) and one proposed model (ACT). In all the models, features fed to a DNN were linguistic features that comprised 675 dimensions and they were normalized to have values ranged from 0.01 to 0.99, and the DNN architectures were feed-forward networks that included 6 hidden layers, each of which has 1024 units and tanh activation functions. The DNN parameters were randomly initialized, and stochastic gradient descent (SGD) was used for the optimization algorithm. The learning rate was set to 0.5 in ACT, 0.005 in SP and LogSP, and 0.002 in MCEP.

The output of MCEP is Mel-cepstrum coefficients that consist of 60 dimensions and their dynamic and acceleration coefficients. This dimension is the default value in Merlin[3], a speech synthesis toolkit for neural network-based speech synthesis. The output of SP and LogSP is spectral envelopes that consist of 513 dimensions (16 kHz), or 1025 dimensions (48 kHz). The output of ACT is the activation that comprised 201 dimensions. This is the concatenation of the activation normalized to sum to 1 (200 dim.), and its power (1 dim.). When calculating NMF, the $l2$ norm of each basis is normalized. The number of NMF iteration was set to 1000.

In MCEP and LogSP, output features were normalized to have zero-mean and unit-variance, and their loss functions were MSE. The linear output layer was used in these models. In SP, output features were normalized to have values ranged from 0.01 to 0.99, and KLD was used for its loss function. In this model, the sigmoid output layer was employed similarly to [3]. In ACT, the activation was normalized to sum to 1, and its power was not normalized. Cross-entropy was employed for the loss function of the normalized activation (norm-act), and D-ISD for that of the power coefficient. The output layer for norm-act was softmax, whereas softplus was used for its power.

We also performed an experiment of bandwidth expansion. By using 50 sentences of parallel speech data (16 kHz and 48 kHz), we constructed parallel dictionaries. These 50 sentences are included in the 450 sentences in the training set. By multiplying the 48 kHz dictionaries and the activation predicted by the DNN trained with 16 kHz speech data, we obtained 48 kHz speech samples (*16k_to_48k*). We compared it with the speech samples predicted by the DNN trained with 48 kHz training data (*48k*).

In all the models, to synthesize test speech samples, F_0 and band aperiodicity measures were extracted from the reference samples, and only the parameters of spectral envelopes were predicted by the DNNs. In MCEP, maximum likelihood parameter generation (MLPG) was utilized in order to consider the dynamic and acceleration coefficients [20].

To objectively evaluate the performance of the models, Mel-cepstral distortion (MCD) was used [20]. MCD is calculated as:

$$\mathrm{MCD[dB]} = \frac{10}{\ln 10}\sqrt{2\sum_{d=1}^{24}\left(mc_d - \hat{mc}_d\right)^2} \qquad (12)$$

where mc_d and $\hat{mc}_d$ is the d-th dimension of the Mel-cepstral coefficient.

For the subjective evaluation, preference AB tests on speech quality were performed. In each test, 10 pairs of utterances from the two focused models were randomly selected, and they were evaluated by 25 participants.

[2] http://hts.sp.nitech.ac.jp/
[3] http://www.cstr.ed.ac.uk/projects/merlin/

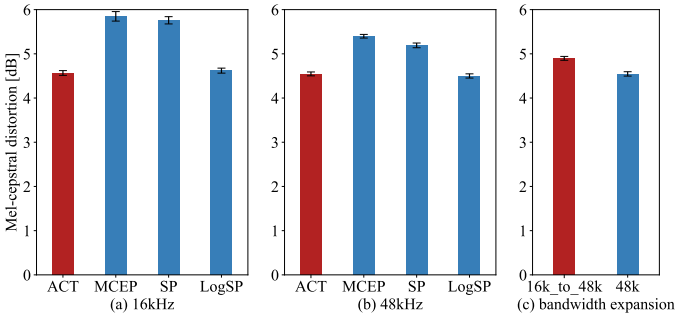

Fig. 3. Mel-cepstral distortion; (a) in 16 kHz, (b) in 48 kHz sampling rates among the four models, and (c) of the experiment of bandwidth expansion. Error bars indicate 95 % confidence intervals.

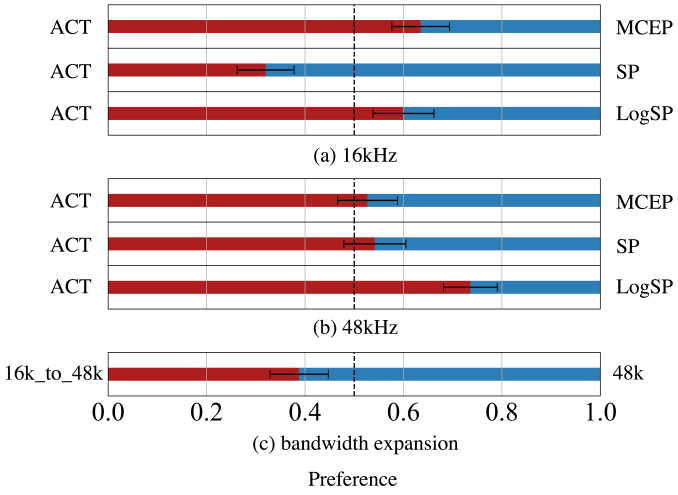

Fig. 4. Subjective results; (a) in 16 kHz, (b) in 48 kHz sampling rates among the four models, and (c) of the experiment of bandwidth expansion. Error bars indicate 95 % confidence intervals.

B. Experimental Results

Figure 3 shows the averaged Mel-cepstral distortions of test utterances in different sampling rates; (a) 16 kHz and (b) 48 kHz among the four models. From Figure 3, the proposed model (ACT) achieved a lower distortion than MCEP and SP in both 16 kHz and 48 kHz sampling rates. In MCEP, the MSE of the features was adopted for the loss function for training of DNN, i.e. DNN was trained to directly minimize MCD. Nevertheless, the proposed model outperformed the MCEP model in the MCD criterion. This would be because NMF can keep the finer structure of spectral envelopes than MCEP. Compared with LogSP, ACT achieved comparable results in the MCD criterion. Although the dimension of the acoustic features in ACT is lower than that in LogSP, ACT can efficiently model the spectral envelopes.

In Figure 4, the results of the subjective evaluations in the different sampling rates are shown. Except for SP in 16 kHz sampling rate, ACT outperformed the other models. It shows that the proposed method can produce more natural synthetic speech especially in wide-band conditions.

The results of the experiment of bandwidth expansion are also depicted in Figures 3 (c) and 4 (c). *48k* achieved the slightly better results than *16k_to_48k*. In *48k*, 450 utterances of wide-band samples were used for training, while *16k_to_48k* utilized only 50 wide-band utterances. Although the acoustic model of *16k_to_48k* is trained by utilizing only narrow-band samples, the proposed scheme of band expansion achieved the reasonable performance, which is slightly worse than the model trained with only wide-band samples.

V. Conclusions

In this paper, we have proposed a combination of NMF with DNN-based speech synthesis. In the proposed method, activity patterns derived from NMF are employed for acoustic features, and they are modeled by DNN. To make it possible for DNNs to model the sparse activity patterns appropriately, we have employed a suitable loss function based on the cross-entropy and the dual Itakura-Saito divergence, which can be derived from the definition of KLD. In addition, the proposed framework has been combined with bandwidth expansion. By preparing a small amount of wide-band samples, the proposed method can generate the wide-band synthetic speech even when the acoustic model is constructed by narrow-band samples. Experimental evaluations show that the proposed method can generate more natural spectral parameters especially in 48 kHz sampling rate. In the experiment of 16 kHz-to-48 kHz bandwidth expansion, wide-band speech samples with the reasonable quality are obtained by the acoustic models trained with narrow-band samples.

As further works, combinations of other NMF applications such as voice conversion and noise reduction with the proposed framework should be investigated. The concept of NMF-based voice conversion would connect the proposed framework with multi-speaker training, and that of NMF-based noise reduction makes it possible to utilize the noisy samples for TTS. Improvement of the quality of the bandwidth expansion and effect of sparseness constraint in NMF need to be considered.

Acknowledgments

This research and development work was supported by the MIC/SCOPE #182103104.

REFERENCES

[1] Keiichi Tokuda, Yoshihiko Nankaku, Tomoki Toda, Heiga Zen, Junichi Yamagishi, and Keiichiro Oura. Speech synthesis based on hidden markov models. *Proceedings of the IEEE*, 101(5):1234–1252, 2013.

[2] Heiga Zen, Andrew Senior, and Mike Schuster. Statistical parametric speech synthesis using deep neural networks. In *2013 IEEE international conference on acoustics, speech and signal processing*, pages 7962–7966, 2013.

[3] Shinji Takaki, Hirokazu Kameoka, and Junichi Yamagishi. Direct modeling of frequency spectra and waveform generation based on phase recovery for dnn-based speech synthesis. In *INTERSPEECH*, pages 1128–1132, 2017.

[4] Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, et al. Tacotron: Towards end-to-end speech synthesis. *arXiv preprint arXiv:1703.10135*, 2017.

[5] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. WaveNet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 2016.

[6] Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, et al. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4779–4783. IEEE, 2018.

[7] Yuchen Fan, Yao Qian, Feng-Long Xie, and Frank K Soong. Tts synthesis with bidirectional lstm based recurrent neural networks. In *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[8] Heiga Zen and Haşim Sak. Unidirectional long short-term memory recurrent neural network with recurrent output layer for low-latency speech synthesis. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4470–4474. IEEE, 2015.

[9] Manu Airaksinen, Bajibabu Bollepalli, Lauri Juvela, Zhizheng Wu, Simon King, and Paavo Alku. Glottdnn-a full-band glottal vocoder for statistical parametric speech synthesis. In *Interspeech*, pages 2473–2477, 2016.

[10] Eunwoo Song, Kyungguen Byun, and Hong-Goo Kang. Excitnet vocoder: A neural excitation model for parametric speech synthesis systems. *arXiv preprint arXiv:1811.04769*, 2018.

[11] Zhen-Hua Ling, Li Deng, and Dong Yu. Modeling spectral envelopes using restricted Boltzmann machines and deep belief networks for statistical parametric speech synthesis. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(10):2129–2139, 2013.

[12] Shinji Takaki, SangJin Kim, Junichi Yamagishi, and JongJin Kim. Multiple feed-forward deep neural networks for statistical parametric speech synthesis. In *Interspeech*, pages 2242–2246, 2015.

[13] Daniel D Lee and H Sebastian Seung. Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems*, pages 556–562, 2001.

[14] Zhizheng Wu, Tuomas Virtanen, Tomi Kinnunen, Eng Siong Chng, and Haizhou Li. Exemplar-based voice conversion using non-negative spectrogram deconvolution. In *ISCA Workshop on Speech Synthesis, SSW8*, pages 201–206, 2013.

[15] Kevin W Wilson, Bhiksha Raj, Paris Smaragdis, and Ajay Divakaran. Speech denoising using nonnegative matrix factorization with priors. In *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4029–4032. IEEE, 2008.

[16] Paris Smaragdis and Bhiksha Raj. Example-driven bandwidth expansion. In *2007 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 135–138, 2007.

[17] Patrik O Hoyer. Non-negative matrix factorization with sparseness constraints. *Journal of machine learning research*, 5(Nov):1457–1469, 2004.

[18] Akira Kurematsu, Kazuya Takeda, Yoshinori Sagisaka, Shigeru Katagiri, Hisao Kuwabara, and Kiyohiro Shikano. ATR japanese speech database as a tool of speech recognition and synthesis. *Speech communication*, 9(4):357–363, 1990.

[19] Masanori Morise, Fumiya Yokomori, and Kenji Ozawa. WORLD: a vocoder-based high-quality speech synthesis system for real-time applications. *IEICE TRANSACTIONS on Information and Systems*, 99(7):1877–1884, 2016.

[20] Tomoki Toda, Alan W Black, and Keiichi Tokuda. Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(8):2222–2235, 2007.