Audio Engineering Society

Convention Paper 9400

Presented at the 139th Convention
2015 October 29–November 1 New York, USA

Extension of Monaural to Stereophonic Sound Based on Deep Neural Networks

Chan Jun Chun[1], Seok Hee Jeong[1], Su Yeon Park[1], and Hong Kook Kim[1,2]

[1]School of Information and Communications
Gwangju Institute of Science and Technology (GIST), Gwangju 500-712, Korea
{cjchun, jeongsh, stellasy0213, hongkook}@gist.ac.kr

[2]Dept. of Electrical and Computer Engineering, City University of New York, NY 10031, USA

ABSTRACT

In this paper, we propose a method of extending monaural into stereophonic sound based on deep neural networks (DNNs). First, it is assumed that monaural signals are the mid signals for the extended stereo signals. In addition, the residual signals are obtained by performing the linear prediction (LP) analysis. The LP coefficients of monaural signals are converted into the line spectral frequency (LSF) coefficients. After that, the LSF coefficients are taken as the DNN features, and the features of the side signals are estimated from those of the mid signals. The performance of the proposed method is evaluated using a log spectral distortion (LSD) measure and a multiple stimuli with a hidden reference and anchor (MUSHRA) test. It is shown from the performance comparison that the proposed method provides lower LSD and higher MUSHRA score than a conventional method using hidden Markov models (HMM).

1. INTRODUCTION

Stereophonic sounds obviously provide a more natural listening experience than monaural sounds [1]. Even though the reproduction systems mostly support the stereo channels, monaural contents still exist since they involve simple and inexpensive recording.

In order to utilize monaural audio content in stereophonic reproduction systems, numerous methods have been proposed [2-5]. By applying two all-pass filters with different phase information to monaural signals, stereophonic signals could be generated easily [2]. In [3], monaural audio was converted using inter-channel coherence (ICC) without additional information. However, as the correlation property varies significantly over time, it is difficult to generate natural sound with this method. As an alternative, Gaussian mixture model (GMM) or hidden Markov model (HMM)-based extension methods have also been proposed [4,5]. Comparing the performance between GMM and HMM-based method, the latter could handle the mismatch problem in the energy trajectory between adjacent audio frames, thus it provided better performance than the GMM-based one [5].

Recently, the performance of deep neural network (DNN)-based algorithms has surpassed that of the

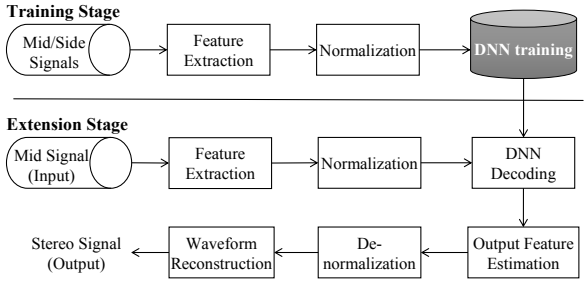

Fig. 1: Block diagram of the proposed mono-to-stereo extension method.

HMM-based method in speech recognition and speech synthesis [6]. Furthermore, DNN-based speech enhancement significantly improved performance compared to the HMM-based method [7], where DNN-based speech enhancement models were trained with features from pairs of noisy and clean speech using minimum mean squared error (MMSE)-based fine-tuning.

Accordingly, a DNN-based extension method that converts monaural to stereophonic sound is proposed. In particular, it is assumed here that monaural signals are the mid signals for the extended stereo signals. After that, the proposed method is used to estimate the side signals from the mid signals by using line spectral frequencies (LSFs) as the DNN features.

The remainder of this paper is organized as follows. Following this introduction, Section 2 proposes the DNN-based mono-to-stereo extension method. After that, Section 3 evaluates the performance of the proposed method by using log spectral distortion (LSD) and a multiple stimuli with a hidden reference and anchor (MUSHRA) test [8]. Finally, this paper is concluded in Section 4.

2. PROPOSED MONO-TO-STEREO EXTENSION METHOD

Fig. 1 illustrates a block diagram of the proposed mono-to-stereo extension method. In the proposed method, monaural signals are assumed to be mid signals for the extended stereo signals, and the side signals are estimated from the mid signals. Here, the mid and side signals, $x_{m,i}(n)$ and $x_{s,i}(n)$, are defined as

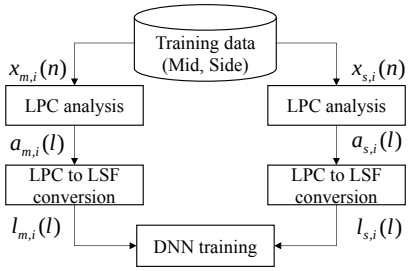

Fig. 2: Procedure of DNN training for the proposed mono-to-stereo extension method.

$$x_{m,i}(n) = (x_{L,i}(n) + x_{R,i}(n))/2, \\ x_{s,i}(n) = (x_{L,i}(n) - x_{R,i}(n))/2 \quad (1)$$

where $x_{L,i}(n)$ and $x_{R,i}(n)$ are the left and right channel signals in the training data at the *i*-th frame, respectively. Next, the residual signals are obtained by performing the *M*-th order linear prediction (LP) analysis [9], such as

$$r_{m,i}(n) = x_{m,i}(n) - \sum_{l=1}^{M-1} a_{m,i}(l) x_{m,i}(n-l) \quad (2)$$

where $a_{m,i}(l)$ and $r_{m,i}(n)$ are the LP coefficients of the mid signals and the residual signals, respectively. In this paper, we set *M* and the frame size as 30 and 1024, respectively. Then, the LP coefficients of the mid signals are converted into the LSF coefficients [10], as illustrated in Fig. 2. Note that the LSF coefficients are widely used for the vocal track parameters as an alternative of LP coefficients [10]. By utilizing the LSF coefficients for the mid signals as the DNN features, the features for the side signals are estimated, as described in the following subsections.

2.1. DNN training stage

For unsupervised pre-training, we first try to train a deep generative model using mid signals by stacking multiple restricted Boltzmann machines (RBMs) [11]. One visible layer of linear variables connected to a hidden layer is a Gaussian–Bernoulli RBM. Moreover, a pile of Bernoulli–Bernoulli RBMs is stacked behind the Gaussian–Bernoulli RBM. In this paper, the number of RBMs was set to 3, 4 and 5. In addition, the number of hidden units was set to 2048 and 4096. The learning rate and splice parameter were set to 0.0005 and 5, respectively.

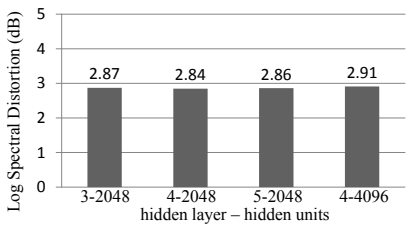

Fig. 3: Comparison of the LSD depending on the number of hidden layers and hidden units used for DNN training.

For supervised fine-tuning, the backpropagation method with the MMSE cost function between the target, referred to the features of the side signals, and the features of the side signals is used [7]. In contrast to pre-training for initializing the parameters in the first hidden layer, the fine-tuning performs supervised training of all the parameters in the networks. The learning rate and the number of iterations were set to 0.008 and 100, respectively.

In order to train the DNN, stereo signals sampled at 32 kHz and 3.5 hours were prepared. These were excerpted from speech and music genres. In order to find the optimal parameters, the LSD between the original and the estimated stereophonic signals was measured along all the frames, where the LSD was defined as

$$d_{LSD} = \frac{1}{KN}\sum_{i=0}^{N-1}\sum_{c=L,R}\sum_{k=0}^{K-1}\left(10\log_{10}\left|\frac{X_{c,i}(k)}{\hat{X}_{c,i}(k)}\right|\right)^2 \text{ (dB)} \qquad (3)$$

where N is the total number of frames. In addition, $X_{c,i}(k)$ and $\hat{X}_{c,i}(k)$ were obtained by applying an K-point discrete Fourier transform and they were the k-th spectral components of the original and the estimated signal at the i-th frame for left or right channel, respectively. Fig. 3 shows the result of the LSD measurements depending on the number of hidden layers and hidden units. As shown in the figure, the LSD was the lowest when the number of hidden layers and hidden units were 4 and 2048, respectively.

2.2. DNN extension stage

In order to estimate the side signals, the estimated LSF coefficients from DNN are converted into the LP coefficients. Then, the estimated side signal, $\hat{x}_{s,i}(n)$, is reconstructed by using the residual signals for the mid signals and the estimated LP coefficients, such as

$$\hat{x}_{s,i}(n) = r_{m,i}(n) + \sum_{l=1}^{M-1}\hat{a}_{s,i}(l)r_{m,i}(n-l) \qquad (4)$$

where $\hat{a}_{s,i}(l)$ is the l-th LP coefficient estimated from the DNN model. Here, the estimated side signals are filtered with a high-pass filter having a cut-off frequency of 150 Hz. Finally, by adding and subtracting the mid and side signals, the stereophonic signals are obtained.

TABLE I
Comparison of the LSD (in dB) between the conventional and proposed extension method

Conventional (HMM-based)	Proposed (DNN-based)
3.04	2.54

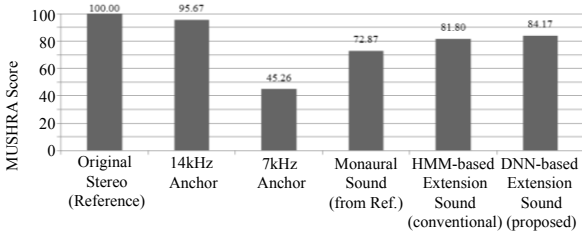

Fig. 4: Comparison of the MUSHRA scores.

3. PERFORMANCE EVALUATION

In order to demonstrate the effectiveness of the proposed method, both objective and subjective tests were carried out. To this end, four audio clips consisting of speech and music signals were taken from the sound quality assessment material (SQAM) database [12]. Note that there was no overlap between the data used for the DNN training and those used for this evaluation. Some speech samples can be found in http://hucom.gist.ac.kr/139AES2015/sample.html. Since these samples were recorded in stereo at a sampling rate of 44.1 kHz, they were down-mixed into monaural signals and down-sampled to 32 kHz. In order to compare the performance of the proposed method with that of the conventional method, the audio clips from the HMM-based extension method [5] were also obtained.

For the objective test, the LSDs between the original stereophonic signals and the corresponding estimated stereophonic signals were measured for two different

extension methods. It was shown from Table 1 that the proposed extension method had a lower LSD value than the conventional HMM-based extension method.

Next, a MUSHRA test [8] was carried out, where two anchors with cut-off frequencies of 7 and 14 kHz were additionally prepared. Eight people with no auditory diseases participated in this test. Each participant was presented with the four stimuli and was asked to rate the stereo quality with a score between 0 and 100. Fig. 4 shows the MUSHRA test result. Each column corresponds to the opinion scores averaged over eight listeners and four audio clips. As shown in the figure, the extended audio from the proposed method achieved an average score of 84.17, which was higher than the down-mixed monaural audio and that of the HMM-based stereo extension method.

4. CONCLUSION

In this paper, we proposed an extension method that converts monaural sound to stereophonic sound using DNNs. To this end, we trained DNN using LSF coefficients of monaural signals to estimate those of stereo signals. We carried out an objective and a subjective test using LSD measure and MUSHRA test, respectively, to compare the performance of the proposed method with that of an HMM-based extension method. It was shown from performance comparison that the proposed extension method could provide lower LSDs and better subjective quality than the HMM-based extension method.

5. ACKNOWLEDGEMENTS

This work was supported in part by the National Research Foundation of Korea (NRF) grant funded by the government of Korea (MSIP) (No. 2015R1A2A1A05001687), and by the ICT R&D program of MSIP/IITP [R0126151034000200 3, Development of hybrid audio contents production and representation technology for supporting channel and object based audio].

6. REFERENCES

[1] F. Rumsey, *Spatial Audio*, Focal Press, Woburn, MA (2001).

[2] M. Schroeder, "An artificial stereophonic effect obtained from a single audio signal," *Journal of the Audio Engineering Society*, vol. 6, no. 2, pp. 74-79 (1958).

[3] N. I. Park and H. K. Kim, "Artificial stereo extension of speech based on inter-channel coherence," *Advanced Science and Technology Letters (ASTL)*, vol. 14, pp. 168-171 (2012).

[4] N. I. Park, K. M. Jeon, C. J. Chun, and H. K. Kim, "Artificial stereo extension based on Gaussian mixture model," *134th AES Convention*, Preprint 8877 (2013).

[5] N. I. Park, K. M. Jeon, S. H. Choi, and H. K. Kim, "Artificial stereo extension based on hidden Markov model for the incorporation of non-stationary energy trajectory," *135th AES Convention*, Preprint 8980 (2013).

[6] G. Hinton, L. Deng, D. Yu, A.-R. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82-97 (2012).

[7] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal Processing Letters*, vol. 21, no. 1, pp. 65-68 (2014).

[8] ITU-R BS 1534, *Method for Subjective Assessment of Intermediate Quality Level of Coding Systems* (2001).

[9] A. Spanias, T. Painter, and V. Atti, *Audio Signal Processing and Coding*, John & Wiley & Sons, Inc., Hoboken, NJ (2007).

[10] X. Mei and S. Sun, "An efficient method to compute LSFs from LPC coefficients," in *Proc. 5th International Conference on Signal Processing*, vol. 2, pp. 655-658 (2000).

[11] Y. Bengio, "Learning deep architectures for AI," *Foundations and Trends(R) in Machine Learning*, vol. 2, no. 1, pp. 1-127 (2009).

[12] EBU Technical Document 3253, *Sound Quality Assessment Material Recordings for Subjective Tests - Users' Handbook for the EBU-SQAM Compact Disc* (1988).