

ACOUSTICAL LETTER

Long short-term memory recurrent-neural-network-based bandwidth extension for automatic speech recognition

Yuuki Tachioka* and Jun Ishii†

Information Technology R&D Center, Mitsubishi Electric Corporation, 5–1–1 Ofuna, Kamakura, 247–8501 Japan

(*Received 31 May 2016, Accepted for publication 27 June 2016*)

Keywords: Speech recognition, Bandwidth extension, LSTM-RNN
PACS number: 43.60.Cg, 43.60.Dh, 43.72.Ne [doi:10.1250/ast.37.319]

1. Introduction

Broad-band speech improves the performance of automatic speech recognition (ASR), but the performance is significantly degraded when broad-band speech is used for training acoustic models and narrow-band speech is input for ASR decoding. Many bandwidth extension (BWE) methods have been proposed for improving a perceptual subjective impression. One of the most effective BWE methods is a Gaussian mixture model (GMM)-based BWE [1]. On the other hand, recently, neural network (NN)-based signal restoration methods have been widely used. Recurrent structures are effective for speech enhancement and, in particular, a long short-term memory recurrent neural network (LSTM-RNN) [2] has high reconstruction performance for signal restoration. In this letter, we propose to use the LSTM-RNN for BWE, and its performance is evaluated for the TIMIT phoneme recognition task.

2. GMM-based BWE

In the field of voice conversion, where the speech of one speaker is converted into that of another speaker, a GMM-based voice conversion technique has been proposed [3]. This type of GMM-based voice conversion is applied to a BWE task [1]. We use this method as a baseline. In the context of voice conversion, a narrow-band speech is the original speech and a broad-band speech is the converted speech. Full covariance GMMs are used for modeling concatenated feature vectors before and after BWE, as shown in Fig. 1. Converted, i.e., BWE, speech is estimated on the basis of the maximum likelihood criteria.

3. LSTM-RNN-based BWE

Generally, for time-series signals, RNNs have higher performance than simple NNs because recurrent structures can consider time-series information. The LSTM-RNN has been proposed to relax the influence of vanishing gradient problems in the RNN and to deal with longer contexts [2]. Its effectiveness has been shown for a speech enhancement task [4]. The LSTM-RNN with a narrow-band speech input and broad-band speech output is trained by an error back-propagation method based on a least-square-error criterion, as shown in Fig. 2.

4. Phoneme recognition experiments on the TIMIT corpus

4.1. Experimental setups

For BWE, two types of speech features were extracted: 1) mel cepstrum (mcep), which is widely used for speech synthesis, and 2) mel-frequency cepstrum coefficient (MFCC), which is widely used for ASR. In the case of a GMM-based BWE, 1) the dimensions of mcep features were 17 for 8 kHz and 25 for 16 kHz, and a total of 84-dimensional features in conjunction with their Δ features were used. For ASR, MFCC features were extracted after signal waves in the time domain were restored from mcep features. 2) The dimension of MFCC features was 13 for both 8 and 16 kHz, and a total of 52-dimensional vectors were used with their Δ features. In this case, the obtained MFCC features were directly input for ASR. SPTK toolkit (ver. 3.7)[a] was used.

In the case of an LSTM-RNN-based BWE, 1) the LSTM-RNN was trained to predict 25-dimensional mcep static features with 25-dimensional Δ, i.e., 50-dimensional in total, features for 16 kHz from the 17-dimensional static mcep with 17-dimensional Δ, i.e., 34-dimensional in total, mcep features for 8 kHz. 2) For MFCC, 26-dimensional MFCC features comprising 13-dimensional static features and Δ features, were used. The "currennt" toolkit (ver.0.2)[b] [4] was used.

The training data of the BWE model and ASR acoustic model were the same as the training data of the TIMIT phoneme recognition task, which was one of the most standard corpora for English ASR. Their performances were evaluated using the development set and evaluation set of the TIMIT in terms of phoneme error rate (PER) using the Kaldi toolkit[c] [5]. For ASR, maximum-likelihood GMM acoustic models were used with MFCC$+\Delta+\Delta^2$ features. To improve the ASR performance, two types of advanced ASR techniques were used. The first one was feature transformation by linear discriminant analysis (LDA) [6] and maximum-likelihood linear transformation (MLLT) [7]; the second one was speaker adaptation by feature-space maximum-likelihood linear regression (fMLLR) [8].

*e-mail: Tachioka.Yuki@eb.MitsubishiElectric.co.jp
†e-mail: Ishii.Jun@ab.MitsubishiElectric.co.jp

[a]http://sp-tk.sourceforge.net/
[b]https://sourceforge.net/projects/currennt/
[c]http://kaldi.sourceforge.net/

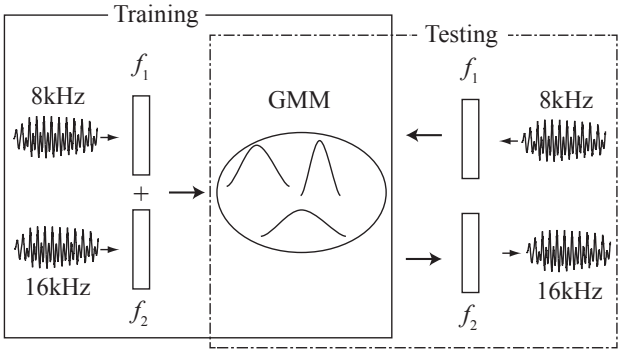

Fig. 1 GMM-based BWE.

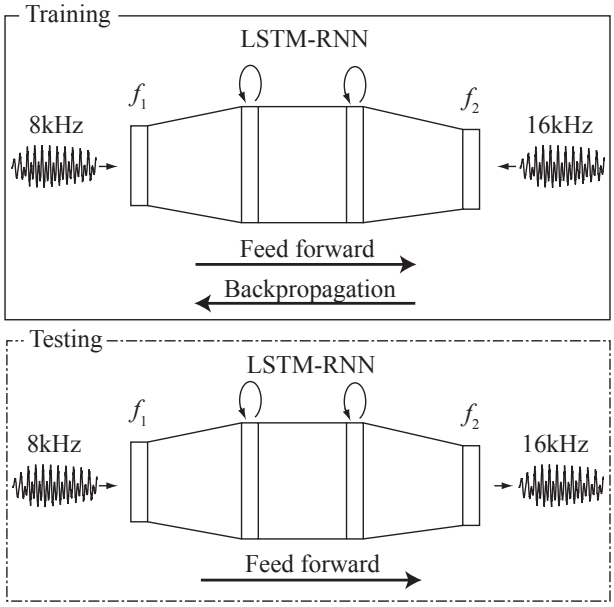

Fig. 2 LSTM-RNN-based BWE.

Table 1 Phoneme error rate (PER) [%] on the **dev** set of the TIMIT phoneme recognition task, evaluating 16 kHz and 8 kHz speech. 8 kHz speech was recognized using 16 kHz and 8 kHz models. MFCC features were used for ASR with advanced ASR techniques such as feature transformation (LDA+MLLT) and speaker adaptation (fMLLR).

		ASR feature		
eval	train	MFCC	+LDA+MLLT	+fMLLR
16k	16k	23.1	21.3	18.9
8k	16k	32.3	28.8	23.4
8k	8k	23.5	22.5	20.3

4.2. Results and discussion

Table 1 shows the baseline performance on the development set. The performance was the highest for the matched case of recognizing 16 kHz speech with 16 kHz models. The second best one was the matched case of recognizing 8 kHz speech with 8 kHz models. These matched cases showed much better performance than the mismatched case of recognizing 8 kHz speech with 16 kHz models. Speaker adaptation decreased the performance gaps between matched and mismatched conditions. When sampling frequencies are different between the training and the test speech, speaker adaptation compensates for the influence of mismatch to some extent but the recognition performance was significantly worse than those of matched conditions.

Table 2 PER [%] on the **dev** set, evaluating GMM- and LSTM-RNN-based BWE (8 kHz → 16 kHz). Mel-cepstrum features were used for BWE. Both gender-dependent (gd) and gender-independent (gi) models were constructed.

	ASR feature					
	MFCC		+LDA+MLLT		+fMLLR	
gd/gi	gd	gi	gd	gi	gd	gi
GMM	28.9	29.3	27.7	27.9	25.2	25.4
LSTM	25.5	25.5	23.8	23.9	22.0	22.1

Table 3 PER [%] on the **dev** set. MFCC features without and with mean normalization (Mean norm.) were used for BWE.

	ASR feature					
	MFCC		+LDA+MLLT		+fMLLR	
Mean norm.	—	✓	—	✓	—	✓
GMM (gi)	36.0	30.4	35.3	29.0	31.9	24.7
LSTM (gi)	25.5	24.7	24.1	23.0	21.5	20.7

Table 2 shows the performance after BWE for mcep features. Both gender-dependent and gender-independent BWE models were prepared, but their performance differences were small for both the GMM and LSTM-RNN cases. For all cases, the LSTM-RNN outperformed the GMM. This shows the effectiveness of a LSTM-RNN-based BWE, as in the case of speech enhancement.

Table 3 shows the performance of directly predicted MFCC features. Gender-independent models were used in the experiments below. There are two cases: without and with mean normalization of input features to the GMM or LSTM-RNN. Mean normalization was essential for the GMM and effective for LSTM. In the two cases of the GMM without speaker adaptation, the performance was degraded, but in the other cases, direct estimation of MFCC improved the performance compared with that of the mcep-based BWE. For the purpose of ASR, a direct estimation of the features suitable for ASR was effective. BWE improved the performance of ASR for 8 kHz speech without switching acoustic models.

Table 4 shows the results of the test set. The tendencies were similar to those of the development set. The LSTM-RNN outperformed the GMM. LSTM using MFCC features achieved the best performance, where the differences between matched cases and BWE cases were less than 1%.

Table 4 PER [%] on the **test** set. Mel-cepstrum features (mcep) and MFCC features were used for BWE.

		ASR feature					
		MFCC		+LDA+MLLT		+fMLLR	
eval	train	Baseline					
16k	16k	24.9		22.3		19.9	
8k	16k	34.8		30.4		25.3	
8k	8k	25.1		23.5		21.0	
		BWE					
BWE feature		mcep	MFCC	mcep	MFCC	mcep	MFCC
GMM (gi)		31.4	32.6	29.8	29.9	26.6	26.1
LSTM (gi)		27.2	25.9	25.2	24.0	23.4	21.9

There is an advantage of the proposed method compared with the use of the matched 8 kHz acoustic model. The proposed method does not require an acoustic model change; thus, it can be widely used for various existing ASR systems without troublesome acoustic model training. If matched acoustic models are needed, training for both 16 kHz and 8 kHz is needed. The training time of acoustic models doubles for each ASR system, whereas the training of the proposed BWE model is required only once. Constructing two types of acoustic models for each system is inefficient because 16 kHz speech has recently come more frequent than 8 kHz speech.

5. Conclusion

We proposed the LSTM-RNN-based BWE and compared its performance with that of a conventional GMM-based BWE in an ASR experiment. Experiments using the TIMIT corpus showed that LSTM-RNN-based BWE was more effective than GMM-based BWE and that predicting MFCC features directly was better than predicting mel-cepstrum features for ASR purposes. The LSTM-RNN achieved a performance equivalent to those of matched cases without the need to switch acoustic models.

References

[1] Y. Wang, S. Zhao, Y. Yu and J. Kuang, "Speech bandwidth extension based on GMM and clustering method," *Proc. 5th Int. Conf. Communication Systems and Network Technologies* (*CSNT*), pp. 437–441 (2015).

[2] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, **9**, 1735–1780 (1997).

[3] T. Toda, A. W. Black and K. Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," *IEEE Trans. Audio Speech Lang. Process.*, **15**, 2222–2235 (2007).

[4] F. Weninger, J. Geiger, M. Wöllmer, B. Schuller and G. Rigoll, "The Munich feature enhancement approach to the 2nd CHiME challenge using BLSTM recurrent neural networks," *Proc. 2nd CHiME Workshop Machine Listening in Multisource Environments*, pp. 86–90 (2013).

[5] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, M. Petr, Y. Qian, P. Schwarz, J. Silovský, G. Stemmer and K. Veselý, "The Kaldi speech recognition toolkit," *Proc. ASRU*, pp. 1–4 (2011).

[6] R. Haeb-Umbach and H. Ney, "Linear discriminant analysis for improved large vocabulary continuous speech recognition," *Proc. ICASSP 92*, pp. 13–16 (1992).

[7] R. Gopinath, "Maximum likelihood modeling with Gaussian distributions for classification," *Proc. ICASSP 98*, pp. 661–664 (1998).

[8] M. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Comput. Speech Lang.*, **12**, 75–98 (1998).