

INVESTIGATION ON NEURAL BANDWIDTH EXTENSION OF TELEPHONE SPEECH FOR IMPROVED SPEAKER RECOGNITION

Phani Sankar Nidadavolu, Vicente Iglesias, Jesús Villalba, Najim Dehak

Center for Language and Speech Processing, Johns Hopkins University, Baltimore, MD

{snidadal, viglesil, jvillal17, ndehak3}@jhu.edu

ABSTRACT

We extend our previous work on training mixed-bandwidth (BW) speaker recognition system by predicting missing information in upperband (UB) of upsampled telephone speech. Mixed-BW systems combine speech from narrowband (NB) and wideband (WB) speech corpora by basic upsampling of NB speech with low-pass filter interpolator, resulting in no information loss in the original WB speech. In this work, we explore the usage of a deep residual full-convolutional neural network (CNN) and a bidirectional long short term memory (BLSTM) network along with a previously proposed deep neural network (DNN) for bandwidth extension (BWE) of NB telephone speech. Speaker recognition systems trained with bandwidth extended features improved in performance over mixed-BW and NB baseline systems. In terms of detection cost function (DCF), the CNN-BWE system improved by 10.78% and 15.96% (relative) in the Speakers In The Wild (SITW) eval core and assist-multi-speaker condition respectively w.r.t. the NB baseline; and improved by 3.21% and 4.13% w.r.t. to the mixed-BW baseline.

Index Terms— Bandwidth extension, speaker recognition, deep residual CNN, BLSTM, transfer learning

1. INTRODUCTION

We ¹ consider the scenario of training speaker recognition systems where we have access to labelled data from two different domains—telephone and microphone speech. Telephone speech is sampled at 8KHz whereas the microphone speech is sampled at 16KHz. From now on, we will call 8KHz data as narrowband (NB) and 16KHz data as wideband (WB) and band between 4-8KHz as upperband (UB). The conventional way of combining datasets with different sampling rate is to downsample all of them to the lowest sampling rate—16KHz to 8KHz in our case. This results in loss of information in the upperband (UB) of the 16KHz data. The downsampled microphone speech along with the telephone speech is then used to train the speaker recognition system. This method works fairly well when the evaluation data set

is also telephone speech. However, it is not optimal when the evaluation data is WB since the UB, which is lost, also contains speaker discriminant information. Another option is to train the system only on the available WB corpus avoiding the need for downsampling. But this approach would lead to poor results when the available WB data is scarce. This is a very common scenario in speaker recognition research, driven mainly by NIST speaker recognition evaluations, where NB telephone speech is available in abundance but WB labelled data is scarce.

In [1], we proposed to upsample the NB data using a simple low-pass filter interpolator and then combine it with the WB corpus to train a mixed-BW x-vector [2] speaker recognition system. Hence, no information loss is incurred. However, this basic upsampling does not predict any information in the UB of the NB corpus. In this work, we investigate several neural architectures to predict the missing information in the UB of the upsampled data. We will refer to this process as *Bandwidth extension* (BWE) throughout the paper. We are mainly interested in using this BWE to improve speaker recognition performance in microphone speech. In this paper, we did not focus on actually recovering the waveform and evaluating the quality of the upsampled signals, which would involve to organize human perceptual tests.

There are few works dealing with this problem in the field of automatic speech recognition (ASR). The authors in [3], propose to jointly train a feed forward deep neural network (DNN) for BWE and ASR system on both NB and WB data. Authors in [4, 5] train a DNN on log-power spectrogram (LPS) features for BWE and then train ASR on bandwidth extended NB data. They also showed that predicting the entire WB is better than just predicting the missing UB, which leads to discontinuities between the NB and the UB spectrum. The authors in [6] proposed a mixed-BW training where the NB filter-bank were zero padded to match the dimension of the WB filter-bank when training the ASR system.

In this work, we compare several neural architectures for BWE in terms of speaker recognition performance. We experimented with a feed forward DNN, a deep residual full-CNN and a BLSTM. The CNN architecture was inspired by the work done in [7] for image style transfer and single image super resolution.

¹Both first and second authors contributed equally in this work

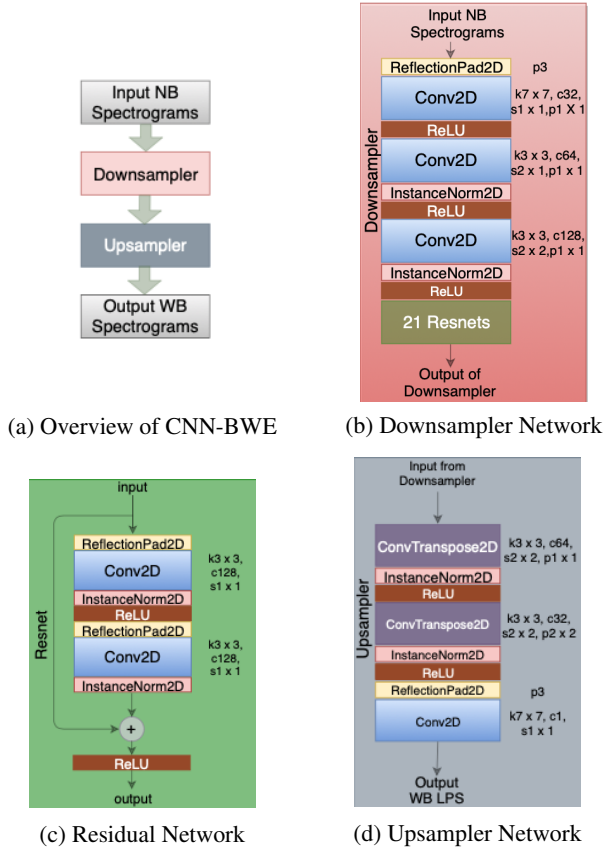


Fig. 1. Individual building blocks of CNN-BWE (k-kernel, s-strides, c-filters, p-padding)

The rest of the paper is organized as follows. In Section 2, we explain the architectures of DNN-BWE, CNN-BWE and BLSTM-BWE systems. Section 3 gives details of datasets used in this work and the training details of those systems. In Section 4, we explain the baseline speaker recognition systems used in this work and the BWE-speaker recognition systems. We conclude with a summary of the work in Section 5.

2. BWE SYSTEMS - ARCHITECTURES

In this section, we describe the network architectures of the BWE systems used in this work.

2.1. DNN-BWE Network

The feed forward DNN that we used for BWE was similar to our previous work [1] and was originally proposed by [5]. The DNN-BWE had 3 hidden layers with 2048 neurons in each hidden layer. 5 past and future frames were appended to the current frame at the input. Rectified linear unit (ReLU) nonlinearity was used in each layer except the output layer where linear activation was used.

2.2. CNN-BWE Network

Our CNN-BWE network architecture was similar to the fully-convolutional network used by [7] for image style transfer and single image super resolution. The network consisted of two major building blocks: a downsampler network and an upsampler network, which essentially work as an encoder-decoder network. An overview of CNN-BWE network is shown in Fig. 1(a). The downsampler network consisted of three 2D convolutional layers followed by several residual blocks [8]. The purpose of the downsampler is to reduce the dimension of the feature maps to a low-dimensional manifold which preserves the properties of speech. It performs in-network downsampling with strides greater than 1, which is computationally efficient since it reduces the number of convolutions by reducing the feature map dimensions. The number of residual blocks was set to 21. For the details of kernel sizes, strides, number of filters and padding dimension in each layer of the downsampler network and the residual network refer to Fig. 1(b) and Fig. 1(c) respectively.

The upsampler network consisted of two deconvolutional layers followed by a convolution layer. The task of the upsampler is to decode the low-dimensional representation into the log-power spectrogram (LPS) of WB speech. The deconvolutional layers increase the feature dimension by a factor of 2 by applying strides of 1/2 while decreasing the number of filters by a factor of 2. For the details of kernel sizes, strides, number of filters and padding dimension in each layer of the upsampler refer to Fig. 1(d).

The entire network consisted of only convolutional and deconvolutional layers. No pooling or fully connected layers were included. Hence the network is termed as “deep residual fully-convolutional neural network”. The fact that it is fully convolutional allows the network to process variable length sequences without increasing the number of parameters.

2.3. BLSTM-BWE Network

We also experimented with a BLSTM with two 512 dimension hidden layers. The non linear activation used was ReLU. The output of the BLSTM was fed to an affine layer, which made the network output dimension to match the required WB features dimension.

3. BWE SYSTEMS - TRAINING SETUP

3.1. Datasets

WB data used in this work comprises of microphone recordings from Mixer6, SRE08 microphone speech and Vox-Celeb [9] data sets. In total, WB data consisted of 30974 utterances from 1871 speakers. All the BWE systems were trained on parallel NB-WB data obtained by downsampling the WB data to 8kHz. The WB microphone data was split into 90%-10% training and validation without speaker overlap.

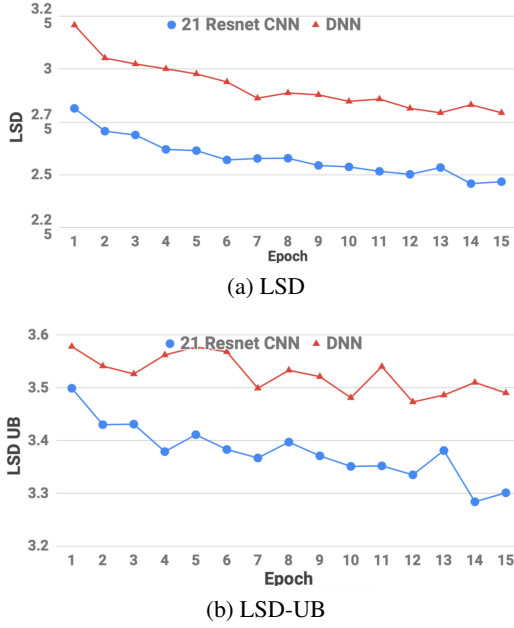


Fig. 2. LSD and LSD-UB comparison of CNN-BWE and DNN-BWE Systems

3.2. Training Details

BWE systems were implemented using PyTorch [10]. As in [4], the BWE systems were trained to predict the entire WB. The input was 129 dimension NB LPS while the output was 257 dimension WB LPS, similar to [4]. Mean squared error (MSE) objective was used. All the networks were trained for a maximum of 50 epochs, where an epoch is completed when we have observed a sample from each utterance. Dropout [11] was used as regularizer (except for CNN-BWE) with 0.4 drop probability.

For the DNN-BWE, for each utterance in the mini-batch we selected 64 frames. We used a context of 5 past and future frames to predict each WB frame. For the CNN-BWE systems, the input was arranged as four dimensional tensor of size $n \times C \times F \times D$, with mini-batch size $n = 1$, channel size $C = 1$, sequence length F and feature dimension $D = 129$. We experimented training with different sequence lengths $F = 129, 257$.

Figure 2 compares the convergence of DNN-BWE and CNN-BWE system trained on sequences of length 129. The performance was measured on the validation data in terms of log-spectral distortion (LSD) and LSD of UB (LSD-UB) at the end of every epoch. LSD measures the reconstruction quality of individual frequencies in LPS domain,

$$\text{LSD}(X, \hat{X}) = \frac{1}{L} \sum_{l=1}^L \sqrt{\frac{1}{K} \sum_{k=1}^K (X(l, k) - \hat{X}(l, k))^2} \quad (1)$$

where X and \hat{X} are the reference and predicted LPS features.

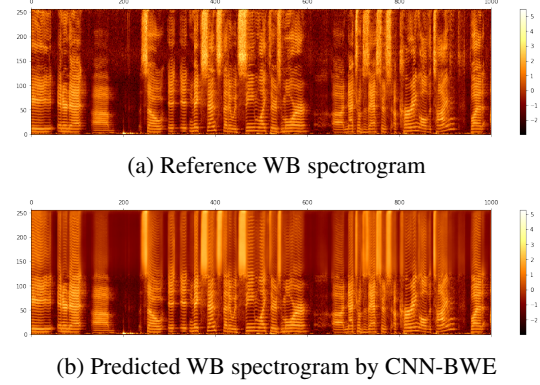


Fig. 3. Comparison of reference and predicted spectrograms by CNN-BWE system.

The CNN-BWE system achieved lower LSD and LSD-UB compared to DNN-BWE. Since we used 21 residual blocks in CNN-BWE network it had, in total, 46 convolutional layers and two deconvolutional layers. In spite of the network depth, the number of parameters of CNN-BWE are less than that of the DNN-BWE. The CNN-BWE network has 6M parameters compared to 16M for DNN-BWE network. Figure 3 shows comparison of spectrograms of original WB utterance from validation set and corresponding predicted WB utterances from a CNN-BWE system.

To compare the performance of CNN-BWE with a recurrent neural network model, we trained a BLSTM-BWE on sequence of lengths of 505 frames. The comparison of BWE systems on speaker recognition performance is given in next section.

4. BWE-SPEAKER RECOGNITION SYSTEMS

We assessed the goodness of the BWE methods for speaker recognition using x-vector speaker recognition systems. Performance was measured on the SITW evaluation set on core-core (single speaker) and assist-multi (multiple speakers in enroll and test) conditions.

4.1. Baseline Systems

We experimented with three baseline x-vector systems: a NB baseline system, a WB baseline system and a mixed-BW baseline system. The NB baseline used 23 MFCC features based on 23 filters Mel filter-bank. The rest of systems used 23 MFCC with 30 filters Mel filter-bank. Features were short-time mean normalized with 3 second sliding window and silence frames were removed. The x-vector system was built using Kaldi [12]. Full rank probabilistic discriminant analysis [13] was used for scoring. PLDA scores were normalized using adaptive symmetric norm (S-Norm) [14].

NB baseline systems represents the conventional way of training speaker recognition systems. Training data comprised of NB telephone data including SRE04-10, Mixer6, Switchboard Phases 1,2 and 3 (91224 utterances from 7001 speakers); and downsampled WB microphone data including Mixer6, SRE08 microphone speech and VoxCeleb (30974 utterances from 1871 speakers). No data augmentation was used.

WB baseline system was trained only on the available WB microphone data. Thus, this system was trained on much less speakers than NB baseline.

The NB baseline system has the disadvantage that the information in the UB of the WB training and evaluation data is lost. No information loss occurs in the WB baseline system but the training data is much smaller, since we have not used NB data. To overcome these two disadvantages, in our previous work [1], we proposed a mixed-BW system where the NB telephone data were upsampled with a low-pass filter interpolator and the WB microphone training and evaluation data were used without any modification. The MFCC configuration used to train this system are similar to the WB baseline system. The mixed-BW system was trained on the same datasets as the NB system.

4.2. BWE-Speaker recognition Systems

The mixed-BW baseline upsamples the NB data but do not estimate the spectrum in the upper-band, which is assumed to be null. The systems in this section used the neural BWE systems to predict the UB spectrum when upsampling. Apart from that, these systems are similar to the mixed-BW baseline. They use the same training datasets and features. The total pipeline for training the BWE speaker recognition systems is as follows. LPS features are extracted from the NB telephone speech and utterance level mean variance normalization is applied. Then, the BWE networks converts NB LPS features into WB LPS features. WB LPS are then converted into MFCC. Finally, the MFCC of the WB data and the BWE MFCC of the NB data are pooled to train the x-vector system.

The experiments with this systems will allow to know whether the BWE networks are able to add speaker discriminant information to the predicted upper-band of the spectrum.

4.3. BWE-Speaker recognition Results

Table 1 summarizes the results of the baseline and BWE-speaker recognition systems. The mixed-BW baseline system performed the best among the baselines. All the BWE systems improved over the baseline systems in terms of both EER and DCF. CNN-BWE system trained on sequences of length 257 performed the best in terms of DCF over all. However, the relative difference in terms of EER between the CNN-BWE and BLSTM-BWE system was not very significant (5.71 and 5.74 respectively). In terms of relative

Table 1. Speaker ID results on SITW (F - sequence length)

	SITW Core			SITW Assist-Multi		
	EER	DCF(1E-2)	DCF(1E-3)	EER	DCF(1E-2)	DCF(1E-3)
Baseline						
NB	6.01	0.5111	0.7105	8.88	0.5569	0.7351
WB	8.02	0.5538	0.7505	8.54	0.5049	0.6880
mixed-BW	5.93	0.4711	0.6713	7.62	0.4890	0.6667
BWE-speaker ID results						
DNN-BWE	5.55	0.4705	0.6516	7.10	0.4812	0.6481
CNN-BWE (F-129)	5.74	0.4737	0.6630	7.08	0.4737	0.6671
CNN-BWE (F-257)	5.71	0.4560	0.6480	7.29	0.4680	0.6475
BLSTM-BWE (F-505)	5.74	0.4621	0.6523	7.35	0.4717	0.6510

improvement of DCF at prior 0.01, the CNN-BWE improved by 10.78% and 15.96% for core and assist-multi-speaker SITW conditions respectively over the NB baseline. Compared to the mixed-BW system, the relative improvements were 3.21% and 4.13% for core and assist-multi-speaker conditions, which justifies our attempt to predict information in the UB for telephone speech for speaker recognition.

5. SUMMARY AND FUTURE WORK

This work investigated the impact of predicting the missing information in the UB of telephone speech on speaker recognition performance. We experimented with a DNN, a deep residual fully-convolutional network (CNN) and a BLSTM network for BWE. Individual x-vector based speaker recognition systems were trained on bandwidth extended features obtained from each of these three systems. The performance of BWE speaker recognition systems was compared to three baseline systems—NB, WB and mixed-BW. All the BWE speaker recognition systems improved in performance compared to all three baselines. The best performer in terms of DCF was the CNN-BWE system trained on sequence lengths of 257. In terms of DCF, the CNN-BWE system showed relative improvement of 10.78% and 15.96% in the SITW eval core and assist-multi-speaker condition respectively w.r.t. the NB baseline; and improved by 3.21% and 4.13% w.r.t. to the mixed-BW baseline, which upsamples NB speech just using a low-pass filter interpolator. In terms of number of parameters, the CNN-BWE model with 21 resnets is the most light weight with $\sim 6M$ parameters compared to DNN-BWE with $\sim 16M$ parameters and BLSTM-BWE with $\sim 18M$ parameters. As future work, we will explore joint model training of BWE system and speaker recognition system with alternate feature representations like filter banks and raw waveform. We will also look into unsupervised BWE using adversarial learning.

6. REFERENCES

- [1] Phani Sankar Nidadavolu, Cheng-I Lai, Jess Villalba, and Najim Dehak, "Investigation on bandwidth extension for speaker recognition," in *Proc. Interspeech 2018*, 2018, pp. 1111–1115.

- [2] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur, “X-vectors: Robust dnn embeddings for speaker recognition,” *Submitted to ICASSP*, 2018.
- [3] Jianqing Gao, Jun Du, Changqing Kong, Huaifang Lu, Enhong Chen, and Chin-Hui Lee, “An experimental study on joint modeling of mixed-bandwidth data via deep neural networks for robust speech recognition,” in *Neural Networks (IJCNN), 2016 International Joint Conference on*. IEEE, 2016, pp. 588–594.
- [4] Kehuang Li, Zhen Huang, Yong Xu, and Chin-Hui Lee, “Dnn-based speech bandwidth expansion and its application to adding high-frequency missing features for automatic speech recognition of narrowband speech,” in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.
- [5] Kehuang Li and Chin-Hui Lee, “A deep neural network approach to speech bandwidth expansion,” in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 4395–4399.
- [6] Jinyu Li, Dong Yu, Jui-Ting Huang, and Yifan Gong, “Improving wideband speech recognition using mixed-bandwidth training data in cd-dnn-hmm,” in *Spoken Language Technology Workshop (SLT), 2012 IEEE*. IEEE, 2012, pp. 131–136.
- [7] Justin Johnson, Alexandre Alahi, and Li Fei-Fei, “Perceptual losses for real-time style transfer and super-resolution,” in *European Conference on Computer Vision*. Springer, 2016, pp. 694–711.
- [8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
- [9] Arsha Nagrani, Joon Son Chung, and Andrew Senior, “Voxceleb: a large-scale speaker identification dataset,” *arXiv preprint arXiv:1706.08612*, 2017.
- [10] Adam Paszke, Sam Gross, Soumith Chintala, and Gregory Chanan, “Pytorch: Tensors and dynamic neural networks in python with strong gpu acceleration,” 2017.
- [11] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov, “Dropout: a simple way to prevent neural networks from overfitting,” *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.
- [12] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al., “The kaldi speech recognition toolkit,” in *IEEE 2011 workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society, 2011, number EPFL-CONF-192584.
- [13] Niko Brümmer and Edward De Villiers, “The speaker partitioning problem.,” in *Odyssey*, 2010, p. 34.
- [14] Niko Brümmer and Albert Strasheim, “Agnitios speaker recognition system for evalita 2009,” in *The 11th Conference of the Italian Association for Artificial Intelligence*. Citeseer, 2009.