Speech Bandwidth Expansion For Speaker Recognition On Telephony Audio

Ganesh Sivaraman, Amruta Vidwans, Elie Khoury

Pindrop, Atlanta, GA, USA

{gsivaraman,avidwans,ekhoury}@pindrop.com

Abstract

Practical applications often require speaker recognition systems to work well for audio files of different sampling rates. However, the performance of speaker recognition systems degrades substantially under a mismatched audio sampling rate between the training and testing conditions. For example, wideband speaker recognition models trained on audio files with a 16kHz sampling rate perform poorly on telephony audio with an 8kHz sampling rate due to the missing higher frequency information. In this paper, we propose a Deep Neural Network (DNN) based system to estimate the speech spectrum in the frequencies above 4kHz for narrowband 8kHz telephony audio. We train the proposed system on speech datasets processed using various simulated telephony codecs. Additionally, we perform speaker recognition experiments by using the bandwidth expansion system as a preprocessor for speaker verification using wideband models. The evaluation datasets used for speaker verification are codec-degraded downsampled Voxceleb1 and SITW, and the NIST SRE 2010 10s-10s condition. We see a significant improvement in the results compared to a simple upsampling with interpolation and low-pass filtering. Additionally, these promising experiments show that the proposed bandwidth expansion system can be used successfully as a data augmentation for training speaker embedding systems.

1. Introduction

Speech technologies like *automatic speech recognition* (ASR) and *automatic speaker verification* (ASV) are being widely deployed in services like call centers, interactive voice response systems, and virtual personal assistant devices. The ASR and ASV systems are often encountered with audio from multiple devices sampled at different sampling rates. Audio recorded and transmitted modern IoT, home assistant devices, and many VoIP based communication channels are often sampled at 16kHz with a wide bandwidth of 0-8kHz. Traditional telephony audio is band-limited to a frequency range of 0.3-3.4kHz, sampled at 8kHz and also encoded with speech coding algorithms. Speech technologies usually obtain superior performance on wideband audio data because of the additional information available in the higher frequency bands. Studies have shown that performance of speaker recognition systems improves with the inclusion of higher frequency bands [1]. A speaker recognition system trained on wideband speech performs poorly on narrowband audio due to the mismatch in the training and testing condition. The missing higher frequency bands in narrowband speech leads to the degraded performance of wideband trained speaker recognition systems on telephony speech.

The mismatch in sampling frequency between the train and test condition can be addressed by two methods - (1) retraining or adapting the models with narrowband audio data, (2) esti-

mating the higher frequency bands from the narrowband audio. Estimation of the higher frequency bands from the lower frequency bands is known as bandwidth expansion (BWE). In this paper, we propose a system for BWE and investigate the efficacy of the BWE systems for speaker verification.

Since the time when speech coding was developed, extending the bandwidth of narrowband audio by estimating the higher frequency information has been an active area of research. Early approaches to BWE were based on parametric linear estimation, nonlinear spline fitting, codebook based methods, and probabilistic statistical methods like Hidden Markov Models (HMM) and Gaussian Mixture Models (GMM). In one of the earliest works on BWE [2], the authors used linear prediction formulation of speech to breakdown the BWE problem into envelope extension, and the excitation signal estimation. They estimated the envelope spectrum using multidimensional filtering and reconstructed the residual excitation signal using a codebook based mapping between narrowband and wideband excitation signals. In another work [3], Non-negative Matrix Factorization (NMF) based algorithm was proposed for BWE. This algorithm was a speaker-dependent system as was the case with other methods at that time. One of the earliest efforts on BWE for telephony audio [4] proposed a cubic spline based estimation of the expanded envelope spectrum, and spectral folding for the excitation spectrum. The authors consider telephony audio compressed using the AMR narrowband codec [5] and also show their method to be robust to background noises. Similar to their predecessors, most of the early statistical parametric approaches to BWE were speaker-dependent. In the earliest effort using neural networks for BWE [6, 7], the authors used spectral folding, low frequency spectral features, and neural network based spectral shaping to estimate the higher frequency bands for telephony audio. This work was focused on improving the subjective quality of the audio.

More recently, various Deep Neural Network (DNN) based systems have been explored for BWE. In [8, 9] the authors proposed a feed-forward DNN based architecture for BWE. The authors showed that their BWE outputs were not only perceptually preferred, but also that their BWE system reduced the ASR word error rate by 4.6%

Previous work on ASR has shown that mixed bandwidth training of DNN based acoustic models provides a word error rate reduction of 18.4% relative to the wide-band trained models [10]. Mixed bandwidth training has also been shown to significantly reduce the EER of speaker verification systems. DNN based BWE also provides further improvement in performance on mixed bandwidth trained speaker recognition systems. Mixed bandwidth training with BWE while testing on narrowband audio has been shown to provide 16.8% relative reduction in EER compared to the wideband baseline [11].

Recent work on DNN-based BWE using neural networks propose multiple architectures for BWE using feedforward

DNNs, deep residual Convolutional neural networks (CNN), and BLSTM networks [12, 13]. The authors performed speaker verification experiments on the speakers in the wild (SITW) dataset to show that their proposed BWE system provided significant improvements in the speaker verification equal error rates (EER). They trained mixed bandwidth x-vector speaker embedding systems using narrowband and wideband audio and showed that their proposed system provided significant improvements even for the mixed BW systems. While their work showed drastic improvement, they did not test on neither codec-degraded downsampled data nor real telephony data.

Most of the previous work has been focused on narrowband audio obtained by simply downsampling wideband audio. In contrast, our work focus on BWE for telephony audio. Telephony audio is more challenging than simply band-limited and downsampled audio because telephony uses a narrow bandpass filter limiting the frequencies to a range of 0.3-3.4kHz. In addition, quantization of the bandlimited audio using audio codecs further introduces distortions in telephony audio.

In this paper, we propose a novel CNN-DNN based architecture. Our proposed architecture consists of a single convolutional and three feed-forward layers, which have a relatively low-computational cost suitable for real-time applications. To support our experimental study, we simulate various codec distortions by encoding and decoding audio files using three telephony codecs. Our BWE system is trained to map audio bandlimited and degraded by telephony codecs to the wideband frequency spectrum. We train our BWE models on clean read speech datasets. We perform speaker recognition evaluations on three different datasets with large domain mismatch between the BWE training data and the speaker recognition evaluation data. Our proposed BWE expansion system provides a 11.1% relative reduction in EER on the SITW-dev core-core protocol. On the real telephony data from NIST SRE 2010 protocol, our BWE system provides a 4.4% reduction in EER, inspite of the significant mismatch between the BWE training data and the conversational NIST SRE data.

The paper is organized as follows: Section 2 details the various datasets used in this work. The codec distortion method used is also explained in this section. A detailed description of the proposed BWE system is provided in Section 3. Section 4 presents the speaker verification system used in this paper. Section 5 explains the speaker verification experiments performed and the results obtained. Section 6 discusses the conclusions and future work.

2. Datasets

In this paper, we have used several speech datasets for three purposes - (1) training the bandwidth expansion (BWE) system, (2) training the speaker recognition system, and (3) speaker recognition evaluation.

2.1. Datasets for BWE Training

We trained the BWE system using Librispeech [14] and VCTK [15] datasets. We used the 100-hour training set of Librispeech dataset, which is a large corpus of English speech obtained from audiobooks. The 100-hour subset of Librispeech consists of 251 speakers. The VCTK dataset is a corpus of clean speech consisting of 44 hours of audio from 109 native English speakers with various accents. Random selection of 99 out of the 109 speakers was performed from the VCTK dataset for training. Overall our BWE training data consisted of 350 unique

speakers. For cross-validation, we used the dev-clean set of Librispeech and 5 speakers from VCTK. The test set for evaluating the BWE performance consisted of 5 speakers from the VCTK dataset and the test-clean set of Librispeech. The Librispeech and VCTK are wideband audio datasets sampled at 16kHz and 48kHz, respectively. We downsampled (using sox) the whole VCTK dataset to 16kHz sampling rate because, in our paper, we consider 16kHz sampling rate as wideband audio. We generated the narrowband version of the training, development, and test sets by simulating audio codec distortions.

2.1.1. Simulated Telephony Codec Distortions

Most BWE studies use audio data downsampled from 16kHz to lower sampling rates by simply band-limiting the audio and resampling. To the best of our knowledge, this paper is the first of its kind to consider audio codec distortions also as part of the narrowband audio. This factor is essential to consider since, in most of the telephony and audio transmission applications, the audio is not only downsampled but also quantized using an audio compression protocol at the transmitter end. At the receiver end, the received bit-streams are decoded according to the corresponding compression applied to reconstruct the audio. The filtering technique and the quantization scheme from the codec causes distortions in the audio. All the BWE systems in this paper are trained on audio compressed and decoded using one of the 3 audio codecs namely-

- Adaptive Multi-Rate Narrowband (AMR-NB) [5]
- Opus [16]
- SILK [17]

All of these audio codecs were used in the narrowband mode with a sampling rate of 8kHz. The table 1 provides some information about the bitrate and common usage of these audio codecs.

Table 1: Description of audio codecs used in this paper

Codec	Bitrate	Applications
AMR-NB	4.75kbps (M475 mode) 12.2kbps (M122 mode)	Mobile telephony
Opus	8-12 kbps (narrowband)	VOIP (Whatsapp , Playstation4 etc.)
SILK	6-20kbps (narrowband)	VOIP (Skype)

A given wideband audio file sampled at 16kHz was passed through a randomly chosen codec to get the 8kHz sampling rate audio file. Executable binaries of the codecs were used to simulate codec distortion for generating the protocols. The VCTK dataset, Librispeech 100 hour training set, dev-clean, and test-clean sets (of the Librispeech dataset) were thus passed through our codec simulator to create the narrowband telephony data for BWE training.

2.2. Datasets for Speaker Recognition

The speaker recognition systems described in this paper are all trained on the VoxCeleb2 dataset. The VoxCeleb dataset [18] consists of audio-visual clips of speech extracted from interview videos from YouTube. Only the audio portion of the dataset is used for the work done in this paper. The audio from the original Voxceleb2 dataset were all sampled with a sampling frequency higher than 16kHz, and we converted them to 16kHz audio for training our speaker recognition system. The VoxCeleb2 dataset consists of 6114 speakers. The dataset consists of more than

1 million utterances amounting to over 2000 hours of speech. We trained our DNN based speaker embedding system on the Voxceleb2 dataset.

We used the VoxCeleb1-E subset of the Voxceleb1 dataset [19] for speaker recognition evaluations in this paper. We created a narrowband (8kHz sampling rate) copy of the Voxceleb1-E subset with codec distortions as outlined in 2.1.1. The codec distorted 8kHz Voxceleb1-E set was used to perform speaker recognition evaluation to assess the efficacy of the BWE system for speaker recognition.

We also used the evaluation set of the Speakers-in-the-wild (SITW) dataset for evaluating the recognition performance with and without BWE. We again converted the SITW-eval set [20] to 8kHz audio using codec distortions as described in 2.1.1.

In order to test our BWE system on realistic telephony data, we chose the evaluation set of the NIST-SRE 2010 data [21] for speaker recognition evaluation. Table 2 shows the list of datasets used in this work and their purposes.

Table 2: List of datasets used and their purposes

Dataset	Purpose
Librispeech [14]	BWE training & testing
VCTK dataset [15]	BWE training & testing
Voxceleb2 [18]	Speaker embedding Training
Voxceleb1 [19]	Speaker verification
SITW [20]	Speaker verification
NIST-SRE2010 [21]	Speaker verification

3. Bandwidth Expansion System

The BWE system proposed in this paper is a CNN-DNN system consisting of a single 1D CNN layer followed by 3 feed-forward layers. The CNN layer at the input contains 64 filters with a kernel size of 5. The feed-forward layers contain 1024 nodes in each layer. Figure 1 shows the block diagram of the proposed BWE system. The input to the network is narrowband log-spectrograms normalized to zero mean and unit variance (z-normalization) for every utterance. We provide 11 frames of 128 dimensional narrowband log-spectrogram as input. The network predicts the 257-dimensional wideband z-normalized log spectrum of the center frame. The network is trained using a mean-squared error loss function between the estimated and the actual wideband log-spectrum. The network is trained for 30 epochs with the adam optimizer [22]. Early stopping criterion using the Librispeech dev-clean and VCTK development data is used to prevent overfitting.
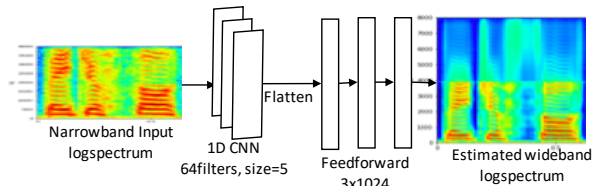

Figure 1: Block diagram of the Bandwidth expansion network.

At inference time, we compute the wideband 257-dimensional spectrogram using simple upsampling of the codec

distorted narrowband input audio file. We use only the lower half of this input spectrogram for predicting the BW expanded spectrum. We z-normalize the 2-D narrowband spectrogram and compose the input to the network by concatenating each frame with 5 frames of context on each side and feed it to the trained BWE system. We use the log-spectral mean and variance statistics of the input narrowband features to denormalize the estimates of the BWE system. Due to the smoothness introduced by the mean-squared-error loss function, the estimates of the BWE system are much smoother than the higher bands of the original 16kHz spectrum. The coarse texture of the higher bands that can be seen in the simple upsampled spectrum of the narrowband input (top panel of figure 3). In order to introduce this texture to the estimates of the higher bands, we add back a fraction (parametrized by α) of the higher bands of the input spectrogram. The parameter α is tuned for the best performance on the speaker recognition task using the SITW development set. Due to the low-pass filtering performed by the codec distortion, the energy of the higher frequency channels is much lower compared to the higher frequency channels of the original wideband spectrum. We compute an average inverse filter by computing the average difference between the wideband log-spectrum and the narrowband log-spectrum on the training dataset. The log-spectrum of the average inverse filter is show in Figure 2. We add this inverse filter (addition in the log-spectrum domain) to the denormalized BWE estimates. The complete denormalization of the BWE estimates can be summarized by the following equation.

$$\log(Y) = (1-\alpha)(\text{BWE}_{\text{pred}} \times \sigma_{\log(X)} + \mu_{\log(X)}) + (\alpha \log(X)) + inv_{filt} \tag{1}$$

where, X is the input log-spectrum of the simple upsampled narrowband audio, Y is the denormalized BW expanded log-spectrum, BWE_{pred} is the output of the BWE system, and inv_{filt} is the inverse filter shown in figure 2. This BW expanded denormalized log-spectrogram is then used to compute the MFCC features for input to the speaker embedding network for speaker verification experiments.
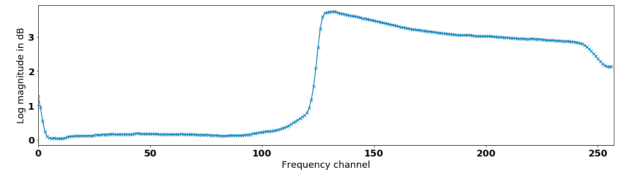

Figure 2: Inverse filter to reverse the codec filtering effect.

Figure 3 shows an example plot of the BW expanded estimate from the BWE system.

We evaluated the performance of the BWE system by computing the Log Spectral Distortion (LSD) between the estimated BW expanded spectrum and the reference wideband spectrum. The LSD measure between a reference spectrogram S, and its estimate $\hat{S}$ is defined as shown in the equation below [23]

$$\text{LSD} = \frac{1}{T}\sum_{t=0}^{T}\sqrt{\frac{1}{k_u - k_l + 1}\sum_{k=k_l}^{k_u}\left[10\log_{10}\left(\frac{|S(t,k)|^2}{|\hat{S}(t,k)|^2}\right)\right]^2} \tag{2}$$

The LSD for the lower frequency bands and the higher frequency bands are reported in table 3 for the test set consisting of Librispeech test set and 5 speakers from the VCTK dataset.
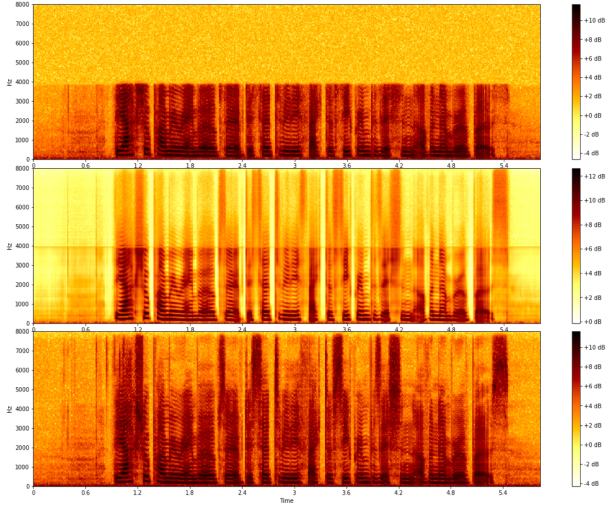
Figure 3: BWE estimate for an example test utterance. The top panel shows the narrowband log-spectrum, the middle panel shows the denormalized BW expanded spectrum estimated by the BWE system, and the bottom panel shows the spectrogram of the original 16kHz audio

The LSD values of the BWE estimates are compared to the LSD between the simple upsampled audio and the reference spectrogram. Note that before computing the LSD, we added the inverse filter showing in figure 2 to the upsampled log-spectrogram. Lower the LSD value, the better is the estimate. Note that for the simple upsampled case, the LSD of the lower frequencies is lower compared to the BWE output. This is because of the distortions introduced by the BWE system in the lower frequencies. In all our speaker recognition experiments, we appended the higher frequency estimates of the BWE system to the lower frequency spectrum of the narrowband audio.

Table 3: LSD results for BWE estimates compared to simple upsampling. LSD_{lf} denotes the LSD computed across the lower half frequency channels (1-128), and LSD_{hf} denotes the LSD computed only on the higher frequency bands (129-257).

Test set	Simple upsampling	BWE system output
LSD_{hf}	1.793	1.291
LSD_{lf}	0.934	1.029

4. Speaker Verification System

Convolution neural network-based speaker embedding systems have outpaced i-vector based systems in all of the speaker verification benchmarks. In this work, we designed a fairly simple CNN architecture.

First, energy-based voice activity detection is applied to get rid of silence segments. Then, 30-dimensional Mel frequency cepstrum coefficients (MFCC) are computed on 20 ms windows with an overlap of 10 ms. Those features are normalized using zero-mean and unit-variance normalization and then fed into the CNN.

The training is done in two steps. The first step consists of training a speaker embedding system using softmax and categorical cross-entropy. The architecture of the network is shown in Figure 4. The network includes five convolutional layers, followed by a statistics pooling layer, two fully-connected layers, and a softmax output layer. The output layer consists of 6,114 target speakers.
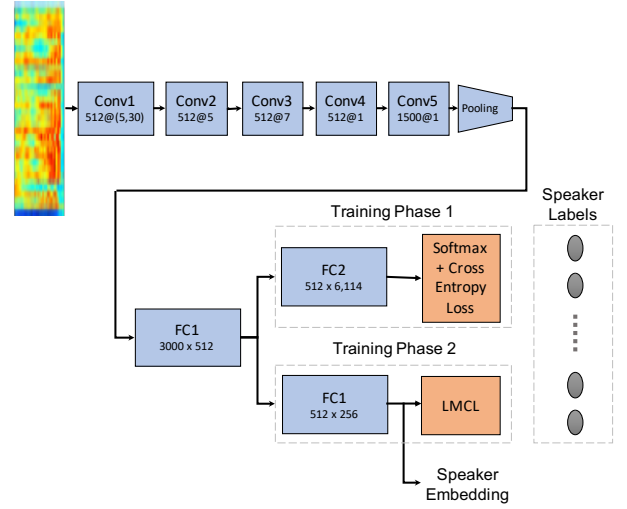

Figure 4: Block diagram of the Speaker embedding network.

The second step consists of removing both the second fully-connected layer and the softmax layer, freezing the remaining layers, and then adding a fully connected layer of size 512 x 256 that is trained using large margin cosine loss (LMCL) [24].

At inference time, the speaker embeddings are extracted from the second fully connected layer. Finally, the scoring is done using a simple *Cosine* metric.

Using the above architecture, two different Speaker embeddings were trained:

- The first is wideband (**WB**) and is trained solely on Voxceleb2 16kHz, without any noise, music, or reverberation augmentation.

- The second is mixed (**WB+BWE**) and is trained on both Voxceleb2 16kHz and the BWE version of the codec degraded version of Voxceleb2.

5. Experiments

The objective of our BWE system is to perform speaker verification on telephony audio. To evaluate the efficacy of the BWE system for speaker verification, we ran experiments on four different protocols: SITW core-core Dev, SITW core-core Eval, Voxceleb1-E, and NIST SRE 2010 10s-10s condition.

SITW core-core Dev is the development set of the core-core protocol of the Speakers in the Wild (SITW) challenge. This is the only set used to tune the α parameter in 1 of the BWE algorithm. For both SITW core-core Dev and Eval, we created a degraded version using the codec simulation method outlined in section 2.1.1.

Voxceleb1-E is a subset of the Voxceleb1 dataset [19]. It consists of 4,715 audio utterances spanning across 40 speakers. The official trial list is balanced with 18,860 positive and 18,860 negative trials. Similar to SITW, we created the codec degraded version of Voxceleb1-E.

The NIST SRE 2010 data [21] consists of conversational telephone speech audio. We used only the 10sec-10sec test protocol where both enrollment and test utterances have only 10

Table 4: Speaker verification results for the WB trained system. Numbers shown are EERs (%).

Condition	SITW-dev core-core	SITW-eval core-core	Voxceleb1-E	NIST-SRE2010 10sec-10sec
16kHz original	6.22%	7.16%	4.12%	N/A
Upsampled	17.54%	18.73%	13.92%	21.25%
BW expanded	15.60%	16.86%	12.33%	20.31%

Table 5: Results for the WB+BWE trained system. Numbers shown are EERs (%).

Condition	SITW-dev core-core	SITW-eval core-core	Voxceleb1-E	NIST-SRE2010 10sec-10sec
16kHz original	5.68%	5.96%	3.82%	NA
Upsampled	10.70%	12.31%	7.79%	16.11%
BW expanded	9.66%	10.63%	6.71%	15.01%

sec of net speech. We avoided using the core conditions to limit the duration mismatch between speaker embedding training and testing.

We performed speaker verification experiments using both speaker embedding systems - WB and WB+BWE that are described in section 4. We ran the experiment on each protocol using both the speaker embedding systems (WB, and WB+BWE). For the SITW-dev, SITW-eval, and Voxceleb1-E protocols we evaluated on three conditions: **16kHz original**, **Upsampled**, and **BW expanded** data. The NIST SRE 2010 protocol being telephony data, we evaluated the Upsampled and BW expanded conditions. In all the experiments, the enrollment and test conditions were matching. For example, in the BW expanded protocol, the enrollment and test utterances were all BW expanded.

Table 4 shows the equal error rates (EERs) computed on the four protocols using the WB speaker embedding system. The first observation is that there is a drastic increase in errors between original 16kHz data and the recovered 16kHz from codec degraded 8kHz data.[1] However, the BW expanded data is less erroneous than simple upsampling technique across all datasets, including real-world telephony data. The relative reduction in EER varies between 4.4% (from 21.25% to 20.31%) on NIST SRE 2010 10s-10s protocol to 11.1% (from 17.54% to 15.60%) on SITW core-core Dev.

In the second set of experiments, we evaluated the WB+BWE speaker embedding system. The results are detailed in Table 5.

It is very encouraging to see that the BWE augmentation helped improving the accuracy on original 16kHz data across all SITW and Voxceleb1 protocols. The drop in EER is 16.8% (from 7.16% to 5.96%) on SITW core-core Eval. Therefore, BWE augmentation can be considered as a new type of data augmentation technique while developing WB speaker recognition system.

More importantly, the drop in EER on recovered 16kHz data is very high, including real-world telephony conditions, where EER on NIST SRE 2010 10s-10s reduces from 21.25% to 16.11% on Upsampled data, and from 20.31% to 15.01% on BW expanded data.

6. Conclusions

In this paper we proposed a BWE system that estimates the higher frequency spectrum of narrowband telephony audio. Our

proposed CNN-DNN architecture for BW expansion consists of 3.13 million parameters which is much lower compared to the sizes of models proposed in recent studies [12]. The BWE system also uses only 5 frames of future context, thus making it suitable for real-time implementation. Our BWE system was targeted to the telephony use case training on speech data degraded by simulated codec distortions.

We performed speaker verification experiments on SITW core-core protocol, Voxceleb-E and the realistic NIST-SRE2010 telephony data. Narrowband audio processed by our BWE system achieved a 11.1% reduction in EER on the SITW-dev set compared to simple upsampled audio. On the NIST-SRE2010 10s-10s protocol which is real telephony conversational speech, our BWE system obtained a relative reduction in EER by 4.4%. Thus, our proposed BWE system improves the speaker verification performance without retraining the speaker embedding.

We also explored the use of the BWE system to augment the training data for the speaker embedding model. Our WB-BWE model trained on wideband and BWE audio provided significant improvements in the accuracy on the original 16kHz data across all SITW and Voxceleb1 protocols. On the SITW-eval core-core protocol, the speaker embedding system augmented with BWE data achieved a 16.8% relative drop (from 7.16% to 5.96%) in EER compared to the system trained only on WB 16kHz data. Hence, the BWE technique can be used as an additional data augmentation technique while developing WB speaker recognition systems.

In the future, we plan to explore further modifications to the target and loss function used to train the BWE system. A loss function explicitly combining the loss due to spectral envelope estimates and the fine structure estimates could improve the estimates of the BWE system. A generative adversarial network style training with adversarial loss from a discriminator system also holds promise for improvements. We plan to further explore speaker discriminative training of the BWE system.

7. Acknowledgments

8. References

[1] S. Hayakawa and F. Itakura, "Text-dependent speaker recognition using the information in the higher frequency band," in *ICASSP*, dec 1994.

[2] Carlos Avendano, Hynek Hermansky, and Eric A Wan, "Beyond NYQUIST: towards the recovery of broadbandwidth speech from narrow-bandwidth speech.," in *EUROSPEECH*, 1995.

[3] Dhananjay Bansal, Bhiksha Raj, and Paris Smaragdis, "Bandwidth expansion of narrowband speech using nonnegative matrix factorization," in *9th European Conference on Speech Communication and Technology*, 2005, pp. 1505–1508.

[4] Laura Laaksonen, Juho Kontio, and Paavo Alku, "Artificial bandwidth expansion method to improve intelligibility and quality of AMR-coded narrowband speech," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2005, vol. I.

[5] Bruno Bessette, Redwan Salami, Roch Lefebvre, Milan Jelínek, Jani Rotola-Pukkila, Janne Vainio, Hannu Mikkola, and Kari Järvinen, "The Adaptive Multirate

---

[1] It is worth noting that this error difference is much more limited when no codec degradation is introduced. But this is not the scope of this study where the focus is on Telephony audio.

Wideband speech codec (AMR-WB)," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 8, pp. 620–636, nov 2002.

[6] Juho Kontio, Laura Laaksonen, and Paavo Alku, "Neural network-based artificial bandwidth expansion of speech," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 3, pp. 873–881, mar 2007.

[7] Hannu Pulakka and Paavo Alku, "Bandwidth Extension of Telephone Speech Using a Neural Network and a Filter Bank Implementation for Highband Mel Spectrum," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 7, pp. 2170–2183, 2011.

[8] 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 *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2015, vol. 2015-Janua, pp. 2578–2582.

[9] Kehuang Li and Chin Hui Lee, "A deep neural network approach to speech bandwidth expansion," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, aug 2015, vol. 2015-Augus, pp. 4395–4399.

[10] Jinyu Li, Dong Yu, Jui Ting Huang, and Yifan Gong, "Improving wideband speech recognition using mixed-bandwidth training data in CD-DNN-HMM," in *2012 IEEE Workshop on Spoken Language Technology, SLT 2012 - Proceedings*, 2012, pp. 131–136.

[11] Hitoshi Yamamoto, Kong Aik Lee, Koji Okabe, and Takafumi Koshinaka, "Speaker Augmentation and Bandwidth Extension for Deep Speaker Embedding," in *Interspeech 2019*. sep 2019, vol. 2019-September, pp. 406–410, ISCA.

[12] Phani Sankar Nidadavolu, Vicente Iglesias, Jesus Villalba, and Najim Dehak, "Investigation on Neural Bandwidth Extension of Telephone Speech for Improved Speaker Recognition," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, may 2019, vol. 2019-May, pp. 6111–6115.

[13] Phani Sankar Nidadavolu, Cheng-I Lai, Jesús Villalba, and Najim Dehak, "Investigation on Bandwidth Extension for Speaker Recognition," in *Interspeech 2018*, ISCA, sep 2018, pp. 1111–1115, ISCA.

[14] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*. apr 2015, vol. 2015-Augus, pp. 5206–5210, IEEE.

[15] Christophe Veaux, Junichi Yamagishi, and Kirsten. MacDonald, "CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit," Tech. Rep., University of Edinburgh. The Centre for Speech Technology Research (CSTR)., Edinburgh, 2016.

[16] Koen Vos, Karsten Vandborg Sørensen, Søren Skak Jensen, and Jean Marc Valin, "Voice coding with opus," in *135th Audio Engineering Society Convention 2013*. 2013, pp. 722–731, Audio Engineering Society.

[17] K Vos, S Jensen, and K Sorensen, "SILK Speech Codec," 2009.

[18] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman, "VoxCeleb2: Deep Speaker Recognition," in *Interspeech 2018*, Hyderabad, India, sep 2018, pp. 1086–1090, ISCA.

[19] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman, "VoxCeleb: A Large-Scale Speaker Identification Dataset," in *Interspeech 2017*, Hyderabad, India, aug 2017, pp. 2616–2620, ISCA.

[20] Mitchell McLaren, Luciana Ferrer, Diego Castan, and Aaron Lawson, "The speakers in the wild (SITW) speaker recognition database," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*. 2016, vol. 08-12-Sept, pp. 818–822, ISCA.

[21] Alvin F. Martin and Craig S. Greenberg, "The NIST 2010 speaker recognition evaluation," in *Proceedings of the 11th Annual Conference of the International Speech Communication Association, INTERSPEECH 2010*, 2010, pp. 2726–2729.

[22] Diederik P. Kingma and Jimmy Lei Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, 2015.

[23] Johannes Abel, Magdalena Kaniewska, Cyril Guillaumé, Wouter Tirry, and Tim Fingscheidt, "An Instrumental Quality Measure for Artificially Bandwidth-Extended Speech Signals," *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 25, no. 2, pp. 384–396, feb 2017.

[24] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu, "Cosface: Large margin cosine loss for deep face recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5265–5274.