DUAL-LABEL DEEP LSTM DEREVERBERATION FOR SPEAKER VERIFICATION

Hao Zhang[1], *Stephen A. Zahorian*[1], *Xiao Chen*[1], *Peter Guzewich*[1], *Xiaoyu Liu*[2]

[1]Department of Electrical and Computer Engineering, Binghamton University, USA
[2]Sony Interactive Entertainment Inc
{hzhang20,zahorian,xchen49, pguzewi1}@binghamton.edu, xiaoyu.liu@sony.com

ABSTRACT

In this paper, we present a reverberation removal approach for speaker verification, utilizing dual-label deep neural networks (DNNs). The networks perform feature mapping between the spectral features of reverberant and clean speech. Long short term memory recurrent neural networks (LSTMs) are trained to map corrupted Mel Filterbank (MFB) features to two sets of labels: i) the clean MFB features, and ii) one-hot vector representing speaker ID, estimated clean pitch tracks, or the fast Fourier transform (FFT) clean spectrogram. The performance of reverberation removal is evaluated by equal error rates (EERs) of speaker verification experiments.

Index Terms— dereverberation, text independent, speaker verification, long short term memory, deep neural networks

1. INTRODUCTION

Speaker verification is the task of determining whether a speaker's claimed identity is true by processing the speech audio. The accuracy of such a task, as well as automatic speech recognition (ASR), suffers when the audio is corrupted by reverberation, which occurs whenever the audio is obtained from a distant speaker [2] [5]. Since reverberant conditions are common, dereverberation methods are of great interest.

One way to reduce degradation caused by reverberation is to map the reverberant speech representation to its clean counterpart, assuming reverberant speech and its corresponding clean version are both available for training. In [9], an algorithm named SPLICE, which is essentially a linear mapping between reverberant and clean speech features, was used for feature compensation. To train a nonlinear mapping or transform, neural networks (NNs) have long been utilized for speech enhancement [12]. With the advent of deep neural networks (DNNs), which are capable of highly comprehensive learning, nonlinear transforms of speech features improved considerably. An example of this approach for ASR is the state-of-the-art acoustic modelling used in [6]. A relatively recent variant of DNNs, deep long short term memory recurrent neural networks (LSTMs), have been reported to give better ASR accuracy than traditional DNNs for feature enhancement [13].

Deep LSTMs are exploited for dereverberation in the present paper and evaluated for the task of speaker verification. We used the bidirectional LSTMs (BLSTMs) structure, as used in [11][13], for its capacity to use both long-term past and future speech feature information to predict clean features for each point in time.

Inspired by the multi-task deep learning research for speech applications [2] where an additional training objective improves the effectiveness of the primary goal, the current study is based on dual-label BLSTMs. The idea is to have two sets of targets during training, so that weights of the network are trained for both sets of targets. The desired goal of the additional targets is guiding and steering the training so that the LSTM processed features will give improved speaker verification results. During training, the inputs are reverberated Mel Filterbank (MFB) outputs and the primary targets are clean MFB outputs, while a one-hot vector representing speaker identity, or a clean pitch track, or a clean fast Fourier transform (FFT) spectrogram, serve as a secondary target. [4] shows a spectrogram-to-MFB DNN mapping outperforms either a spectrogram-to-spectrogram or MFB-to-MFB mapping in terms of robust ASR. The mapping across different frequency domains is also probed in current work. In contrast to [4], the proposed method performs both MFB-to-spectrogram and MFB-to-MFB mappings simultaneously. Using vectors representing speaker identity as secondary labels is a intuitive approach. The YAAPT pitch estimator is used to make the clean pitch targets [14].

2. DUAL-LABEL LSTM NETWORKS

LSTMs were developed as an improved version of Recurrent Neural Networks (RNNs) in that the gradient vanishing problem is mitigated. In the current work, which uses deep LSTMs, a multi-label approach has been implemented and tested using the PyTorch Library [10]. The idea is to have two sets of targets during training, so that weights of the network are optimized for both sets of targets. Many recent studies demonstrate that a related second task can improve the training for the original task, and hence improve its performance [2].

In Figure 1, every blank rectangle block is an LSTM unit. The primary goal is to map the corrupted log scaled MFB outputs to underlying clean ones. In line with the dual-label approach, there is an additional target which can be i) a one-hot vector representing speaker identity, ii) a clean pitch track, or iii) a clean FFT spectrogram. Note that one significant difference between these three additional targets is that the pitch track is one dimensional whereas the one-hot vector has 594 dimensions and the FFT has 100 dimensions. Thus one of these additional targets has much lower dimensionality than the 31-dimensional MFB features and two others have much higher dimensionality. Note that the secondary labels only assist the training of LSTMs, and ultimately the networks are used only to produce dereverberated MFB outputs.

2.1. Network Structure

2.1.1. LSTM specifics

Bidirectional LSTMs were chosen because they can take advantage of long-term both previous and future reverberant speech input to predict the current clean label. The long-term property is consistent with the signal property of reverberation. The dimension of the hidden weights, namely the number of cells, is 256, and the number of hidden layers is 4. Both numbers were heuristically chosen, based on some preliminary experiments. The input size of the network is dictated by the dimension of MFB outputs, i.e., 31.

2.1.2. Loss function

Mean square error (MSE) was chosen as the loss function and the loss on both targets are weighted equally as given in Equation (1). The MSE loss function has the advantage of computational simplicity. With the introduction of secondary targets which have different dimensionality than the inputs, MSE values noticeably decrease during training, which gives feedback about the network performance.

$$Loss = .5 \times Loss(Target1) + .5 \times Loss(Target2) \quad (1)$$

2.1.3. Batch normalization

Batch Normalization was applied right after every LSTM layer to adjust and scale the activations.

2.1.4. Hidden layers for secondary target

In Figure 1, the dashed-line box labeled "Hidden Layer" illustrates that two additional linear layers precede the secondary target, when the secondary target is a one-hot vector or an FFT spectrogram. Because the input MFB features and secondary targets are different in dimensionality, the extra hidden layers are helpful to make the mapping more accurate. The number of hidden layers was heuristically chosen.

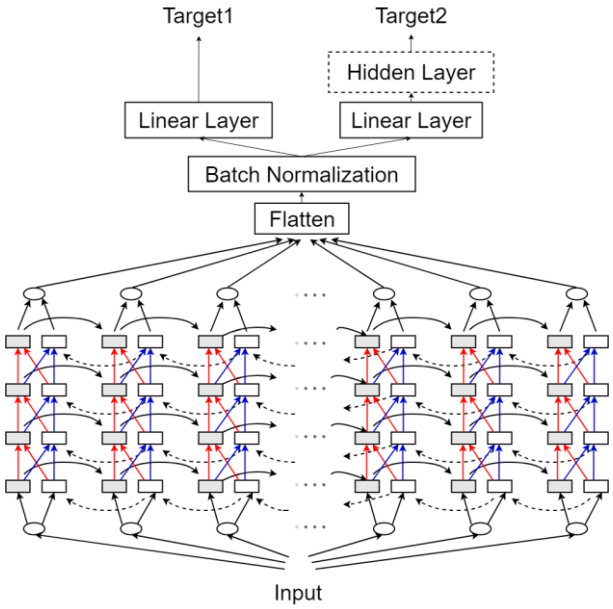

Figure 1. LSTM networks architecture.

2.1.5. Visualization

In Figure 2, (a) shows Mel Filter spectrogram of a 10-second clean utterance, and (b) shows the spectrogram of same utterance artificially reverberated, while (c) is the spectrogram where reverberation is removed by one-label LSTM processing. It can be observed that the LSTM DNN has successfully removed most of the reverberation energy.

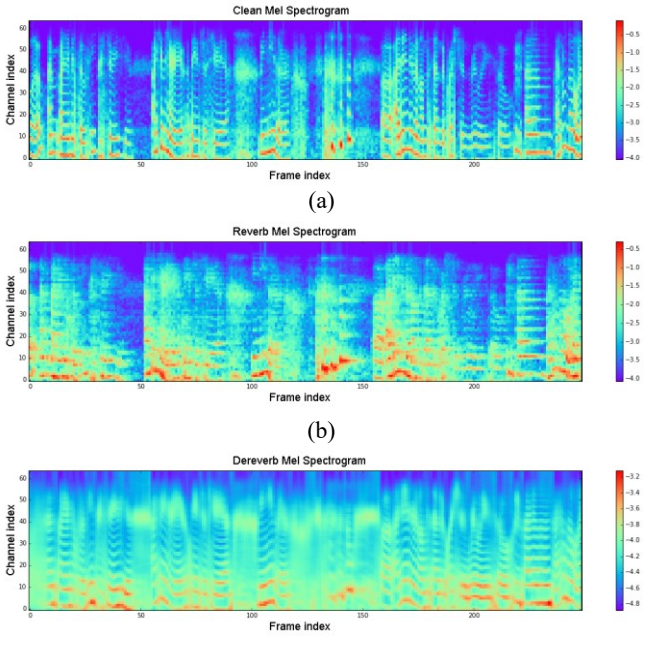

Figure 2. LSTMs dereverberation visualization.

2.2. Inputs and targets of networks

MFB outputs are the inputs and primary targets of networks. Table 1 lists the specifics of the MFB outputs. The same frame length and frame space were used for all features used in the present work.

Table 1. Parameters for MFB features.

Parameter	Value
Frame Length	25 ms
Frame Space	10 ms
Number of Mel Filters	31
Pre-emphasis coefficient	0.97

One-hot vectors representing speaker identities are utilized as secondary targets. Every speaker in the training database is assigned a unique one-hot vector which serves as the secondary target. Every speaker's unique one-hot vector can differentiate the training for each speaker, to some degree. As a result, the trained networks process each speaker's features somewhat differently, which, we hypothesized, could improve speaker verification performance. In our case, one-hot vectors each have 594 elements (the number of speakers in the database), with a 1 value for the ID of a particular speaker, and 0s for the remaining 593 elements.

Pitch tracks extracted from clean speech by the YAAPT pitch estimator [14] are also used as an auxiliary set of labels. We hypothesized that using pitch would improve reverberant to clean mappings in the low frequency range.

FFT spectrograms of clean speech were also used as an auxiliary set of labels. The number of frequency bins at every frame was chosen to be 100 so the spectrogram has reasonably high resolution but does not require excessive computations for the network training. One motivation for choosing a spectrogram as a secondary label is to pursue the performance gain from the cross-frequency-domain mapping [4] and also probe the reason behind the gain.

3. DATABASE

Telephone speech from the Mixer 6 Database [1] was used for speaker verification tests. For this real telephone speech, a caller (channel 1) stays silent approximately half of the time and listens to the other caller (channel 2). To remove the large silent gaps in each telephone channel, voice activity detection (VAD) algorithms were employed. Then 30 seconds of continuous speech were extracted from the beginning of every VAD processed channel, which served as a sample for a speaker. In this way, 8820 equal-length sentences from 594 speakers (302 females and 292 males) were prepared.

For testing, artificial reverberation was added corresponding to the large room, far microphone (T60=0.7s) condition as per the Reverb2014 challenge data [7]. Essentially, clean sentences were convolved with the room impulse response (RIP) from [7].

4. EXPERIMENTS

From the data described in Section 3, 6010 pairs of reverberated and clean waveforms were used for training LSTMs, where clean waveforms were responsible for MFB outputs, pitch tracks and FFT spectrograms. After training, all reverberant waveforms to be used in the speaker verification experiment were passed through the network for processing. Speech from 100 speakers who were not present in the training data were used where 11 sentences per speaker were used for enrollment and 1 sentence per speaker was used for evaluation.

Using the network processed sentences, speech features were computed, consisting of standard 13 Mel-frequency cepstral coefficients (MFCCs), deltas and delta-deltas. As is generally done, the first cepstral coefficient was replaced by energy. There were 39 features total.

Finally, speaker verification experiments were performed using the Alize [8] iVector system which includes a universal background model (UBM) and probabilistic linear discriminant analysis (PLDA) tools. The key parameter settings for Alize are listed in Table 2.

Table 2. Parameters for MFB features.

Parameter	Number
Number of UBM mixtures	1024
iVector dimension	200
PLDA Eigenvoice dimension	100
PLDA Eigenchannel dimension	50

Three cases for each type of feature/dereverberation combination were evaluated. These refer to the data for training, enrollment and testing, as listed below. Except for the baseline experiment, all data are processed by corresponding networks, because in practical applications whether given audio is clean or reverberated is unknown.

- Clean data for training, enrollment and testing (CCC).
- Clean data for training and enrollment while reverberant data for testing (CCR).
- Clean data for training while reverberant data for enrollment and testing (CRR).
- Reverberant data for training, enrollment and testing (RRR).

5. RESULTS AND DISCUSSIONS

Figure 3 illustrates an example of input, target and LSTMs processed MFB features, using same 2-second speech. Frame indices are converted to seconds. Using min-max normalization (2), the magnitude of all features was normalized to a 0 to 1 range.

$$\boldsymbol{A}_{normalized} = \frac{[\boldsymbol{A} - \min(\mathbf{A})]}{[\max(\boldsymbol{A}) - \min(\boldsymbol{A})]} \qquad (2)$$

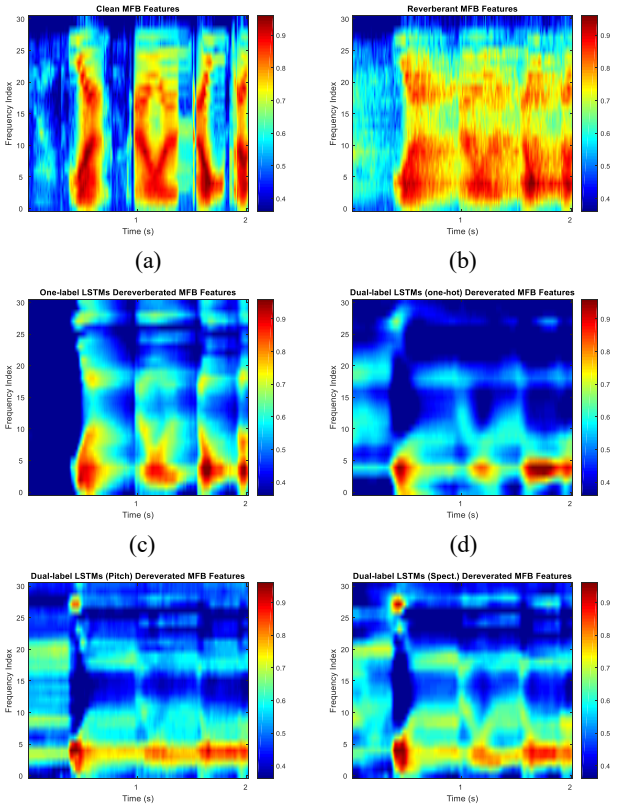

Figure 3. MFB features visualization where (a) is clean, (b) is reverberant, and (c)(d)(e)(f) are LSTM processed.

In Figure 3, six MFB spectrogram figures illustrate the effects of the various LSTM processing configurations used in this study. For reference, panel (a) depicts the MFB spectrogram for clean speech and panel (b) depicts the MFB spectrogram for reverberant speech. Panel (c) depicts the MFB spectrogram for one-label LSTM dereverberated speech. Panels (d), (e), and (f) depict spectrograms for dual-label LSTMs with secondary labels based on speaker ID, pitch, or 100 sample FFTs respectively. The MFB spectrogram is extracted from speech "yeah I am" from Mixer 6 Database.

These plots show that the one-label LSTMs restore most of the energy contours as shown in (c). However, the energy in (c) is overly concentrated on the edges of each segment of speech, while many near-edge regions appear to have too low energy. The fact that high energy regions are very likely to be voiced speech important for automatic speaker verification, and that soft speech is highly distorted by reverberation and thus should deemphasized for speaker verification applications, is a warning that one-label LSTMs may not perform well for preprocessing for SID tasks. Evidently, the energy-level decreases from peaks rather gradually for clean MFB features (a). In (d), features processed by dual-label LSTMs with secondary speaker ID labels have lesser energy than (c). The reason is that the one-hot vectors contain mostly zeros and that LSTMs are trained to reduce energy under the influence of zero-filled secondary targets. Energy contours in (d) are more continuous than in (e) or (f) presumably because the corresponding LSTMs have different secondary labels. Possibly since a pitch track has far fewer dimensions than one-hot vectors or an FFT spectrogram, dereverberated MFB features produced by the pitch version dual-label LSTMs have smoother energy contours. Compared to (f), speech energy in (e) is visibly more concentrated in lower frequencies, presumably because pitch (secondary targets of LSTMs which produced (e)) is of low frequencies.

Table 3. Dual-label LSTMs EERs (%).

EERs of	CCC	CCR	CRR	RRR	AVG
Baseline MFCCs	**2.22**	13.72	11	7.48	8.61
One-label LSTMs	4.97	15	9.06	4.09	8.28
Dual-label LSTMs 1-hot	5.83	**12**	8.7	4.01	7.64
Dual-label LSTMs pitch	4.88	12.87	8.6	**4**	**7.59**
Dual-label LSTMs spect.	5	14	**8.5**	4.12	7.91

In the row called "Dual-label LSTMs 1-hot," the networks have two labels i) MFB outputs from clean speech, and ii) one-hot vectors representing the speakers' identities. The "Dual-label LSTMs pitch" row has the secondary targets of pitch tracks estimated using YAAPT, while the "Dual-label LSTMs spect." row has the secondary targets of clean FFT spectrograms.

As shown in Table 3, the dual-label LSTMs cases outperform their one-label counterpart by small margins, which shows the merits of the dual-label structure. Relative improvement from baseline are as follows (in EERs):

- One-label LSTMs 3.78 % reduction.
- Dual-label LSTMs 1-hot 11.27% reduction.
- Dual-label LSTMs pitch 11.82% reduction.
- Dual-label LSTMs spect. 8.13% reduction.

Thus the lowest EER was obtained using pitch tracks as secondary labels. Introducing the one-hot vectors also substantially improved average EERs. Although spectrograms have higher resolution (much higher dimensionality representing complete spectrum), it is possible that the spectrogram details "misguided" the training process whose objective should be solely dereverberation rather than performing a "Mel-frequency-to-log-frequency" mapping. As a side note, dual-label LSTMs were also tested with a secondary label identical to the primary label, namely MFB features, but no benefit was found.

Follow on work includes testing this general approach with varying degree of reverberation.

6. ACKNOWLEDGEMENTS

This material is based on research sponsored by the Air Force Research Laboratory under contract FA8750-15-C-0266 awarded to Minerva Systems & Technologies, LLC and Binghamton University.

7. REFERENCES

[1] L. Brandschain, D. Graff, and K. Walker. (2013). Mixer 6 Speech LDC2013S03. Hard Drive. *Philadelphia: Linguistic Data Consortium.*

[2] R. Giri, M. L. Seltzer, J. Droppo, and D. Yu, "Improving speech recognition in reverberation using a room-aware deep neural network and multi-task learning," in *Proc. ICASSP. IEEE*, 2015, pp. 5014–5018.

[3] P. Guzewich and S. Zahorian, "Improving Speaker Verification for Reverberant Conditions with Deep Neural Network Dereverberation Processing" in *INTERSPEECH 2017* August 20–24, 2017

[4] K. Han, Y. He, D. Bagchi, E. FoslerLussier, and D. Wang, "Deep neural network based spectral feature mapping for robust speech recognition," in *Proc. Interspeech, 2015*, pp. 2484–2488.

[5] G. Heigold, I. Moreno, S. Bengio, and N. Shazeer, "End-to-end text-dependent speaker verification," in 2016 *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE*, 2016, pp. 5115–5119.

[6] G. Hinton, L. Deng, G.E. Dahl, A. 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, Nov. 2012.

[7] K. Kinoshita, M. Delcroix, S. Gannot,E. Habets, R. Haeb-Umbach, W. Kellermann, V. Leutnant, R. Maas, T. Nakatani, B. Raj, A. Sehr, and T. Yoshioka, "A summary of the REVERB challenge: state-of-the- art and remaining challenges in reverberant speech processing research" *EURASIP Journal on Advances in Signal Processing*, 2016.

[8] A. Larcher, J. Bonastre and H. Li, "ALIZE 3.0 - Open-source platform for speaker recognition," *IEEE SLTC Newsletter*, 2013.

[9] D. Li, A. Acero, M. Plumpe and X. Huang. "Large-vocabulary speech recognition under adverse acoustic environments". *Proc. ICSLP*. 806-809, 2000.

[10] A. Paszke et al., Automatic differentiation in PyTorch, 2017

[11] H. Sak, A. Senior, K. Rao, and F. Beaufays, "Fast and Accurate Recurrent Neural Network Acoustic Models for Speech Recognition," in *INTERSPEECH, 2015.*

[12] E. Wan and A. Nelson, "Networks for speech enhancement," in *Handbook of Neural Networks for Speech Processing*, S. Katagiri, Ed. Norwell, MA, USA: Artech House, 1998

[13] F. Weninger, J. Geiger, M. Wollmer, B. Schuller, and G. Rigoll, ¨ "Feature enhancement by deep LSTM networks for ASR in reverberant multisource environments," *Computer Speech and Language*, vol. 28, no. 4, pp. 888–902, 2014.

[14] S. Zahorian and H. Hu, A spectral/temporal method for robust fundamental frequency tracking, *J. Acous. Soc. Am.* 123(6), 4559–4571, 2008.