Spectral distortion model for training phase-sensitive deep-neural networks for far-field speech recognition

Chanwoo Kim[1], Tara Sainath[2], Arun Narayanan[2], Ananya Misra[2], Rajeev Nongpiur[3], and Michiel Bacchiani[2]

[1]Samsung Research, [2]Google Speech, [3]Nest

[1]chanw.com@samsung.com, [2,3]{tsainath, arunnt, amisra, rnongpiur, michiel}@google.com

Abstract

In this paper, we present an algorithm which introduces phase-perturbation to the training database when training phase-sensitive deep neural-network models. Traditional features such as log-mel or cepstral features do not have have any phase-relevant information. However features such as raw-waveform or complex spectra features contain phase-relevant information. Phase-sensitive features have the advantage of being able to detect differences in time of arrival across different microphone channels or frequency bands. However, compared to magnitude-based features, phase information is more sensitive to various kinds of distortions such as variations in microphone characteristics, reverberation, and so on. For traditional magnitude-based features, it is widely known that adding noise or reverberation, often called Multistyle-TRaining (MTR), improves robustness. In a similar spirit, we propose an algorithm which introduces spectral distortion to make the deep-learning models more robust to phase-distortion. We call this approach Spectral-Distortion TRaining (SDTR). In our experiments using a training set consisting of 22-million utterances with and without MTR, this approach reduces Word Error Rates (WERs) relatively by 3.2 % and 8.48 % respectively on test sets recorded on Google Home.

1 Motivation

1. Recently, phase sensitive features have been introduced (e.g. [1, 2, 3]). which assumes all the microphones are ideal.
2. These features are sensitive to phase distortion arising from microphone distortion, reverberation, auralization, etc in real environments.
3. We may intentionally add phase distortion to the training set so that these features become more robust.

2 Entire Structure (Fig. 1)

1. Room simulator generates simulated far-field utterances.
2. Spectral Distortion Model(SDM) applies magnitude and phase distortion for each utterance.
3. Complex FFT (CFFT) feature is obtained.
4. Factored Complex Linear Prediction (fCLP) mimics the filter-and-sum operation in the spectral domain [1].
5. The output is then passed to a complex linear projection layer [4].
6. The acoustic model pipeline consists of a stack of LSTM layers followed by a DNN (LDNN) layer [5].

3 Spectral Distortion Model (SDM)

The spectrum distortion procedure is summarized by the following pseudo-code (Fig. 2):

```
for each utterance in the training set do
    for each microphone channel of the utterance do
        Create a random Spectral Distortion Model (SDM) using (1).
        Perform Short-Time Fourier Transform (STFT).
        Apply this transfer function to the spectrum.
        Re-synthesize the output microphone-channel using OverLap Addition (OLA).
    end for
end for
```

The Spectral Distortion Model (SDM) is described by the following equation:

$$D_l(e^{j\omega_k}) = e^{am_l(k)+jp_l(k)}, \qquad 0 \le k \le \frac{K}{2}, \quad 0 \le l \le L-1. \tag{1}$$

where l is the microphone channel index and L is the number of microphone channels. In the case of Google Home, since we use two microphones, $L = 2$. a is a scaling coefficient defined by $\ln(10.0)/20.0$. k is the discrete frequency index, ω_k is defined by $\omega_k = \frac{2\pi k}{K}$ where K is the Discrete Fourier Transform(DFT) size. $m_l(k)$ and $p_l(k)$ are Gaussian random samples pulled from the following Gaussian distributions $\mathbf{m}$ and $\mathbf{p}$ respectively:

$$\mathbf{m} \sim \mathcal{N}(0, \sigma_m^2) \tag{2a}$$

$$\mathbf{p} \sim \mathcal{N}(0, \sigma_p^2) \tag{2b}$$

4 Acoustic Model Training

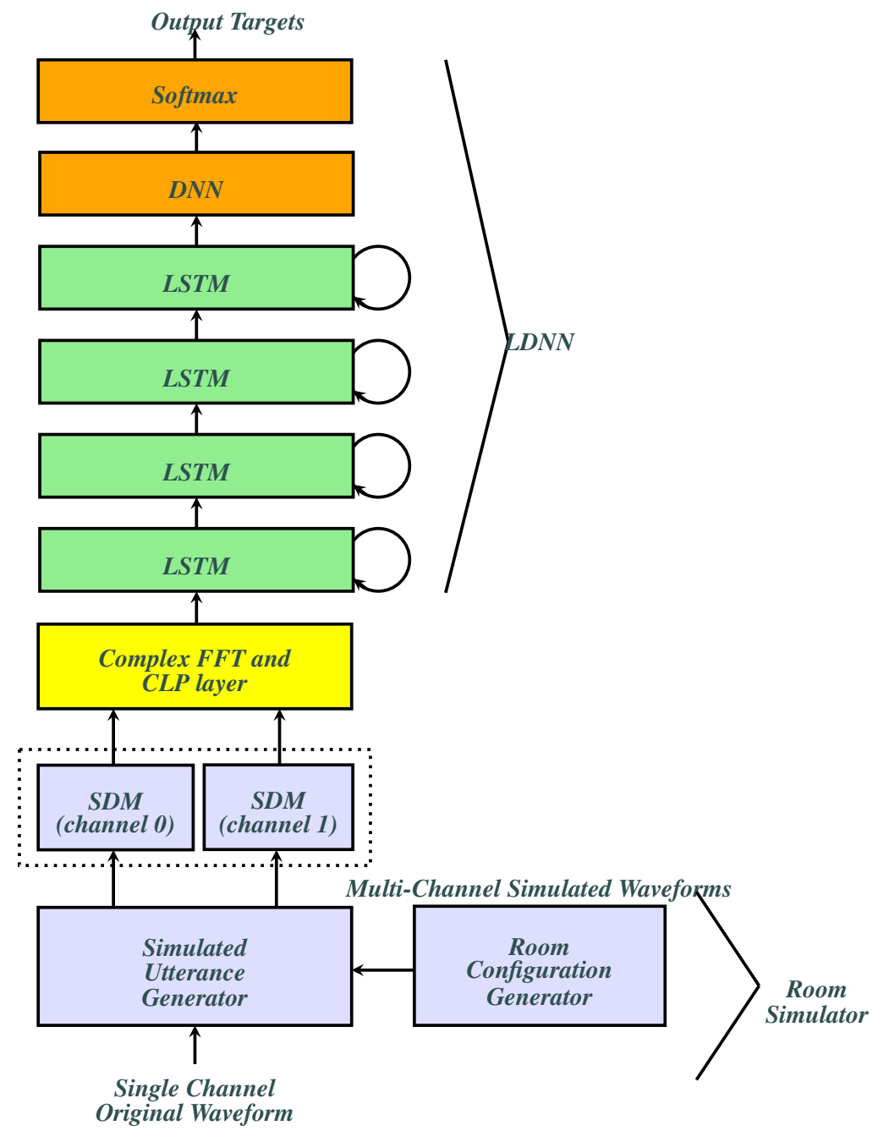
Output Targets
Softmax
DNN
LSTM
LSTM
LSTM
LSTM
LDNN
Complex FFT and CLP layer
SDM (channel 0)
SDM (channel 1)
Multi-Channel Simulated Waveforms
Simulated Utterance Generator
Room Configuration Generator
Room Simulator
Single Channel Original Waveform

Figure 1: The architecture for acoustic model training using the room simulator and LSTMs and a DNN (LDNN) [6, 2].

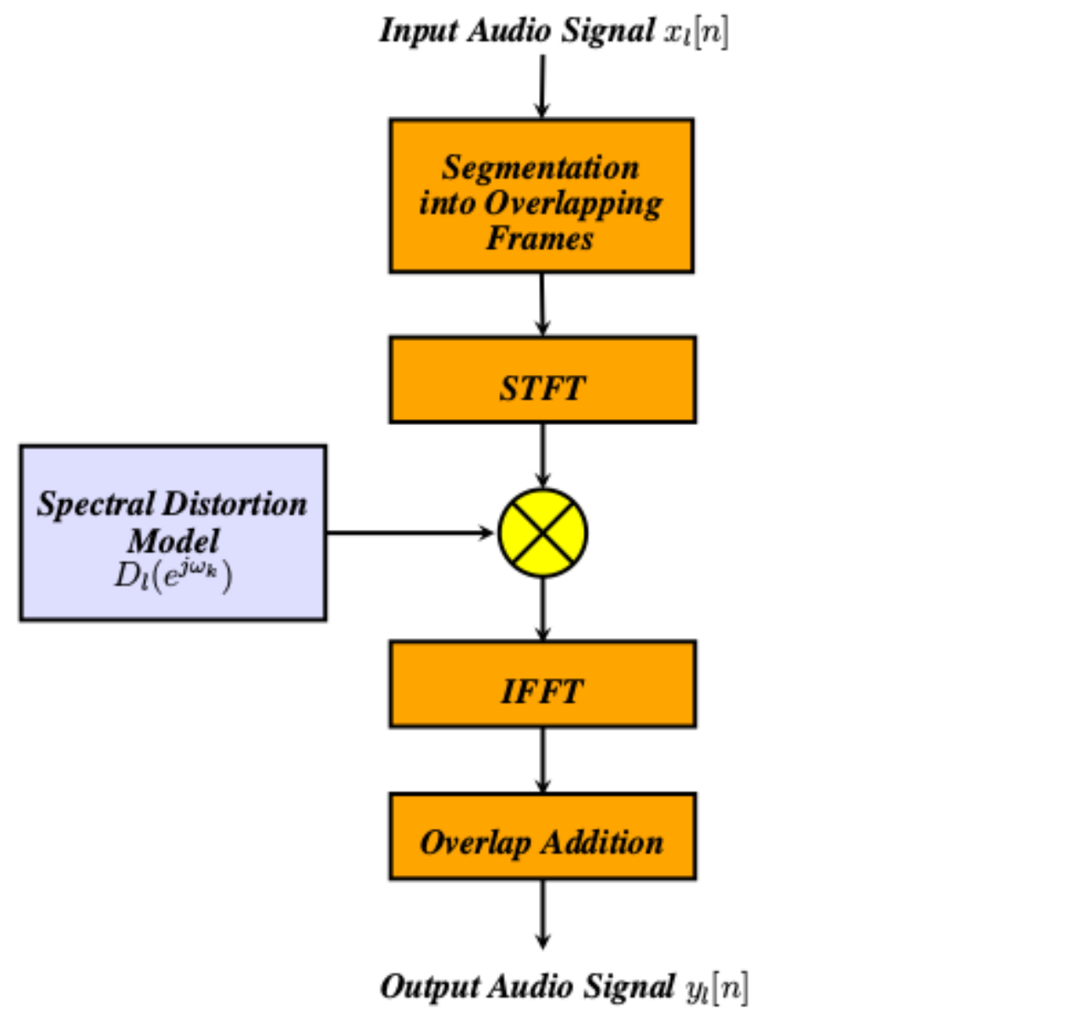
Input Audio Signal $x_l[n]$
Segmentation into Overlapping Frames
STFT
Spectral Distortion Model $D_l(e^{j\omega_k})$
IFFT
Overlap Addition
Output Audio Signal $y_l[n]$

Fig. 2: A diagram showing the structure of applying Spectrum Distortion Model (SDM) in (1) to each microphone channel. Note that l in this diagram denotes the microphone channel index.

5 Experimental Results

- For training, we used an anonymized 22-million English utterances (18,000-hr), which are hand-transcribed.
- For evaluation, we used around 15-hour of utterances (13,795 utterances) obtained from anonymized voice search data. We also generate noisy evaluation sets from this relatively clean voice search data.
- The "room simulator" in [2] was used to generate simulated noisy utterances.
- For Multistyle TRaining, we used the same configuration used in training the Google Home system [2, 1].
- Rerecorded test sets are prepared using three different real Google Home devices.
- MDTR stands for magnitude distortion training without distorting phase.
- PDTR stands for phase distortion training without distorting magnitude.

Table 1: Word Error Rates (WERs) using the PDTR training

	baseline	$\sigma_p = 0.1$	$\sigma_p = 0.4$	$\sigma_p = \infty$
frame length	62.00%			
10 *ms*		57.16 %	56.74 %	54.03 %
32 *ms*		59.03 %	57.14 %	**49.77** %

Table 2: Word Error Rates (WERs) using the MDTR training

	baseline	$\sigma_m = 0.5$	$\sigma_m = 1.0$	$\sigma_m = 2.0$
frame length	62.00%			
10 *ms*		60.39 %		
32 *ms*		**52.21** %	53.03 %	55.37 %

Table 3: Word Error Rates (WERs) using the PDTR and MTR training

	MTR baseline	$\sigma_p = 0.1$	$\sigma_p = 0.4$	$\sigma_p = \infty$
frame length	29.34%			
10 *ms*		28.63 %	**28.40** %	29.78 %
32 *ms*			29.28 %	30.34 %
160 *ms*		28.69 %	31.36 %	37.82 %

Table 4: Word Error Rates (WERs) using the MDTR and MTR training

	MTR baseline	$\sigma_m = 0.5$	$\sigma_m = 1.0$	$\sigma_m = 2.0$
frame length	29.34%			
10 *ms*		31.13 %		
32 *ms*		**28.46** %	28.78 %	28.70 %
160 *ms*			29.01 %	29.55 %

Table 5: Word Error Rates (WERs) obtained with the PDTR ($\sigma_m = 0.0$, $\sigma_p = 0.4$) training

	Baseline	PDTR	Relative improvement (%)
Original Test Set	12.02 %	12.32 %	-2.53 %
Simulated Noise Set 1	20.34 %	20.72 %	-1.86 %
Simulated Noise Set 2	47.88 %	46.69 %	2.50 %
Rerecording using "Device 1"	50.14 %	42.87 %	14.51 %
Rerecording using "Device 2"	48.65 %	43.32 %	10.95 %
Rerecording using "Device 3"	56.27 %	51.30 %	8.83 %
Rerecording with youtube background noise	76.01 %	71.42 %	6.04 %
Rerecording with multiple interfering speaker noise	78.95 %	74.80 %	5.26 %
Average from rerecording sets	**62.00** %	**56.74** %	**8.48** %

Table 6: Word Error Rates (WERs) obtained with the PDTR ($\sigma_m = 0.0$, $\sigma_p = 0.4$) training combined with room-simulator based MTR in [7]

	MTR	PDTR + MTR	Relative improvement (%)
Original Test Set	11.97 %	11.99 %	-0.17 %
Simulated Noise Set 1	14.73 %	15.03 %	-2.04 %
Simulated Noise Set 2	19.55 %	20.29 %	-3.79 %
Rerecording using "Device 1"	21.89 %	20.86 %	4.71 %
Rerecording using "Device 2"	22.23 %	21.29 %	4.22 %
Rerecording using "Device 3"	22.05 %	21.65 %	1.81 %
Rerecording with youtube background Noise	34.83 %	34.21 %	1.78 %
Rerecording with multiple interfering speaker noise	44.79 %	44.00 %	1.76 %
Average from rerecording sets	**29.34** %	**28.40** %	**3.20** %

RMS-PDCW system shows relatively 5.3 percent improvement over the MTR-baseline.

6 Conclusions

In this paper, we described Spectral Distortion TRaining (SDTR) and its subsets Phase Distortion TRaining (PDTR) and Magnitude Distortion TRaining (MDTR). These training approaches apply the Spectral Distortion Model (SDM) to each microphone channel of each training utterance. This algorithm is developed to make the phase-sensitive neural net model robust against various distortions in signals. Our experimental results show that the phase-sensitive neural-net trained with PDTR is much more robust against real-world distortions. The final system shows relatively 3.2 % WER reduction over the MTR training set in [2] for Google Home.

References

[1] B. Li, T. Sainath, A. Narayanan, J. Caroselli, M. Bacchiani, A. Misra, I. Shafran, H. Sak, G. Pundak, K. Chin, K-C Sim, R. Weiss, K. Wilson, E. Variani, C. Kim, O. Siohan, M. Weintraub, E. McDermott, R. Rose, and M. Shannon, "Acoustic modeling for Google Home," in *INTERSPEECH-2017*, Aug. 2017, pp. 399–403.

[2] C. Kim, A. Misra, K.K. Chin, T. Hughes, A. Narayanan, T. Sainath, and M. Bacchiani, "Generation of simulated utterances in virtual rooms to train deep-neural networks for far-field speech recognition in Google Home," in *INTERSPEECH-2017*, Aug. 2017, pp. 379–383.

[3] T. Sainath, R. J. Weiss, K. W. Wilson, B. Li, A. Narayanan, E. Variani, M. Bacchiani, I. Shafran, A. Senior, K. Chin, A. Misra, and C. Kim, "Multichannel signal processing with deep neural networks for automatic speech recognition," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, Feb. 2017.

[4] E. Variani, T. Sainath, I. Shafran, and M. Bacchiani, "Complex Linear Projection (CLP): A Discriminative Approach to Joint Feature Extraction and Acoustic Modeling," in *INTERSPEECH-2016*, Sept. 2016, pp. 808–812.

[5] T. N. Sainath, O. Vinyals, A. Senior, and H. Sak, "Convolutional, long short-term memory, fully connected deep neural networks," in *IEEE Int. Conf. Acoust., Speech and Signal Processing*, Apr. 2015, pp. 4580–4584.

[6] T. Sainath, R. Weiss, K. Wilson, A. Narayanan, and M. Bacchiani, "Factored spatial and spectral multichannel raw waveform CLDNNs," in *IEEE Int. Conf. Acoust., Speech, Signal Processing*, March 2016, pp. 5075–5079.