Power-law nonlinearity with maximally uniform distribution criterion for improved neural network training in automatic speech recognition

Chanwoo Kim, Mehul Kumar, Kwangyoun Kim, and Dhananjaya Gowda

Samsung Research, Seoul, South Korea

{chanw.com, mehul3.kumar, ky85.kim, d.gowda}@samsung.com

1 Summary about Maximum Uniformity of Distribution (MUD)

- We developed a novel approach for obtaining the nonlinearity function in a *data-driven* way.
- This approach is based on the assumption that neural-network training would be easier and the converged parameters would show better performance when feature distribution is not too much skewed or too much concentrated in an extremely narrow interval.
- The criterion is to maximize the Uniformity of Distribution.
- We develop two different types of MUD.
  - Power Function-Based MUD
  - Histogram-Based MUD
- Power function-based MUD shows better results than the histogram-based MUD.
- We achieved 4.02 % WER on test-clean and 13.34 % WER on test-other without using any Language Models (LMs).
- This approach may be extended to other domains as well.

2 Maximization of Distribution Uniformity

- The MUD nonlinearity is applied to mel-filterbank energy for each channel.

2.1 Power-function based maximization of distribution uniformity

- The transformation based on the power function:

$$\mathbf{Y} = \sigma_p(\mathbf{X}) = (X - x_{\text{min}})^{\alpha}. \tag{1}$$

- We expect Y to follow the following uniform distribution:

$$\mathbf{Y} \sim \mathcal{U}(0, (x_{\text{max}} - x_{\text{min}})^{\alpha}). \tag{2}$$

- The log likelihood of the data X assuming the PDF in (2) is given by (Refer to the paper for mathematical details):

$$\begin{aligned}\mathcal{L}(\alpha; X) &= \sum_{i=0}^{N-1} \ln p_{\mathbf{X}}(x_i) \\ &= \sum_{i=0}^{N-1} \ln \left[\frac{\alpha (x_i - x_{\text{min}})^{\alpha-1}}{(x_{\text{max}} - x_{\text{min}})^{\alpha}} \right] \\ &= N \ln(\alpha) + (\alpha - 1) \sum_{i=0}^{N-1} \ln(x_i - x_{\text{min}}) - N\alpha \ln(x_{\text{max}} - x_{\text{min}}). \end{aligned} \tag{3}$$

- The following $\hat{\alpha}$ maximizes the log-likelihood of $\mathcal{L}(\alpha; X)$ (Refer to the paper for mathematical details).

$$\hat{\alpha} = \frac{1}{\ln(\max\{x_i - x_{\text{min}}, \delta\}) - \frac{1}{N}\sum_{i=0}^{N-1} \ln(x_i - x_{\text{min}})}. \tag{4}$$

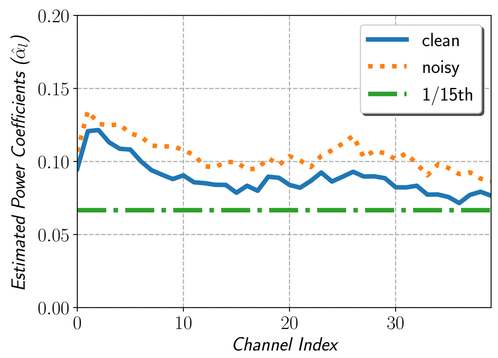
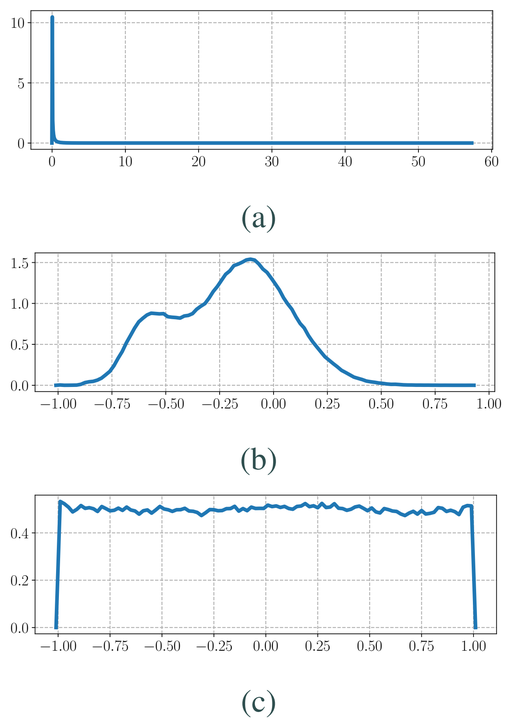


Figure 1: The estimated power coefficients for each mel filterbank channels using (4).

2.2 Histogram-based maximization of distribution uniformity

- The distribution is uniformized using the empirical CDF (Refer to the paper for mathematical details):

$$\mathbf{Y} = \sigma_{np}(\mathbf{X}) = F_u^{-1}\left(\hat{F}_x(\mathbf{X})\right) = \hat{F}_x(\mathbf{X}). \tag{5}$$



(a)

(b)

(c)

Figure 2: The Probability Density Functions for the third filterbank channel of (a): the mel filterbank energy $p[m, l], l = 3$, (b): the power-function based MUD output of this mel filterbank energy in (1), (c) the histogram based-MUD in (5).

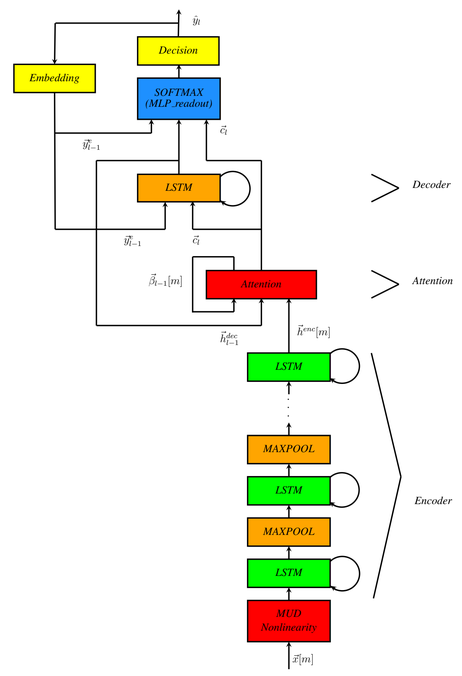


Figure 3: The structure of the entire end-to-end speech recognition system.

3 End-to-End Speech recognition

We used the RETURNN speech recognition system [1] with various modifications:

- MoCha[2] and the modified beam search decoder.
- Gradient clipping, modified learning-late warm-up, and so on.

4 Experimental Results

- **LibriSpeech experiments** (960-hr training and evaluation on LibriSpeech `test-clean` and `test-other`)

Table 1: Word Error Rates (WERs) obtained with MFCC, Power Mel filterbank coefficients, power function-based MUD processing, and histogram-based MUD Processing on the LibriSpeech corpus [3].

Neural Network Structure		MFCC	$(\cdot)^{\frac{1}{15}}$	Power Function-Based MUD	Histogram-Based MUD
1024 cell	test-clean	7.09 %	**7.04** %	7.10 %	7.13 %
ULSTM	test-other	20.60 %	19.76 %	**19.64** %	20.03 %
MoCha	average	13.85 %	13.40 %	**13.37** %	13.58 %
1536 cell	test-clean	4.06 %	**3.94** %	4.02 %	4.11 %
BLSTM	test-other	13.97 %	13.56 %	**13.34** %	14.10 %
Full-Attention	average	9.02 %	8.75 %	**8.68** %	9.11 %

- The Power-Function Based MUD shows slightly better result than the power-law nonlinearity with the power coefficient of $\frac{1}{15}$ [4].
- We hypothesize that the reason why histogram-based MUD does slightly worse than the power- function based MUD is that it somewhat obscured the energy boundary between speech vs non-speech.

References

[1] A. Zeyer, K. Irie, R. Schlüter, and H. Ney, "Improved training of end-to-end attention models for speech recognition," in *INTERSPEECH-2018*, 2018, pp. 7–11. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2018-1616

[2] K. Kim, K. Lee, D. Gowda, J. Park, S. Kim, S. Jin, Y.-Y. Lee, J. Yeo, D. Kim, S. Jung, J. Lee, M. Han, and C. Kim, "attention based on-device streaming speech recognition with large speech corpus," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, Dec. 2019 (accepted).

[3] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *IEEE Int. Conf. Acoust., Speech, and Signal Processing*, April 2015, pp. 5206–5210.

[4] C. Kim and R. M. Stern, "Power-Normalized Cepstral Coefficients (PNCC) for Robust Speech Recognition," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, pp. 1315–1329, July 2016.