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 N. Gowda

Samsung Research

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

email@address

Abstract

In this paper, we describe the Maximum Uniformity of Distribution (MUD) algorithm with the power-law nonlinearity. In this approach, we hypothesize that neural network training will be easier and the converged parameters will show better performance if features are more uniformly distributed. We propose two different types of MUD approaches: power function-based MUD and histogram-based MUD. In these approaches, we first obtain the Mel filterbank coefficients. For each filterbank channel, we apply nonlinearity functions. With the power function-based MUD, we apply a power-function based nonlinearity where the power function coefficient is chosen to maximize the likelihood assuming that the nonlinearity output follows the uniform distribution. With the histogram-based MUD, the empirical Cumulative Density Function (CDF) from the training database is employed to transform the original distribution into a uniform distribution. In MUD processing, we do not use any prior knowledge (e.g. logarithmic relation) about the energy of the incoming signal and the perceived intensity by a human. Experimental results using an end-to-end speech recognition system demonstrate that Power-function based MUD shows better result than the conventional Mel Filterbank Cepstral Coefficients (MFCCs). On the Librispeech database, we could achieve 4.02 % WER on `test-clean` and 13.34 % WER on `test-other` without using any Language Models (LMs). To the best of our knowledge, this is the best result on these databases without using an LM.

Index Terms: Deep-Neural Network Model, end-to-end speech recognition, feature distribution, nonlinearity function, power function

1. Introduction

After the breakthrough of deep learning technology [1, 2, 3, 4], speech recognition accuracy has improved dramatically. Recently, speech recognition systems are widely used not only in smart phones and Personal Computers (PCs) but also in standalone devices in far-field environments. Examples include voice assistant systems such as Amazon Alexa and Google Home [5, 6].

Using the capabilities of neural networks, researchers have explored raw-waveform features [7, 8] or complex Fast Fourier Transform (FFT) features [6, 5]. However, log-mel filterbank coefficients or Mel Filterbank Cepstral Coefficients (MFCCs) [9] still remains the dominant form as features of the automatic speech recognition systems [10, 11, 12]. This is because the conventional features are simpler than the neural network-based features such as raw-waveform features [13] while showing good performance. In log-mel and MFCC, the log-nonlinearity is employed to represent the relationship between the perceived sound intensity by a human and the filterbank energy [14]. In more recent features such as Power Normalized Cepstral Coefficients (PNCCs), the power-law nonlinearity with the power coefficient of $\frac{1}{15}$ is employed. In our previous study [15, 16, 17, 18] this power-law nonlinearity has been shown to be more robust against additive noise. Both the log-law nonlinearity and the power-law nonlinearity with this specific coefficient of $\frac{1}{15}$ were motivated by the rate intensity relation of the human auditory system.

In this paper, we take a completely different approach. Instead of trying to model the human auditory system directly, we try to find a nonlinearity function which maximizes the uniformity of distribution. We refer this approach to Maximum Uniformity of Distribution (MUD) approach. This approach is based on the assumption that even though neural networks have remarkable capabilities in classifying input features, training would be easier and the converged parameters would show better performance if feature distribution is not too much skewed and features are not too much concentrated in an extremely narrow interval. It has been known that the distribution of mel filterbank energy is very sharp and skewed [19, 17]. Thus, it is usually not possible to use mel filterbank energy as features without using any compressive nonlinearity. We proposed two different types of MUD approaches: power function-based MUD and histogram-based MUD. In these approaches, we first obtain the Mel filterbank energy. With the power function-based MUD, we apply a power-function based nonlinearity where the power function coefficient is chosen to maximize the likelihood assuming that the nonlinearity output is the uniform distribution. With the histogram-based MUD, the empirical Cumulative Density Function (CDF) is obtained from the training database to transform the original distribution into uniform distribution. In these two approaches, we do not use any prior knowledge about the relationship between the energy of the incoming signal and the perceived intensity by a human. However, as will be discussed in Sec. , the nonlinearities obtained using the MUD assumption is surprisingly similar to those obtained from human auditory systems. Experimental results with an end-to-end speech recognition system demonstrate that Power-function based MUD shows better result than the conventional Mel Filterbank Cepstral Coefficients (MFCCs) while Histogram-based MUD shows comparable results to the MFCC processing.

The rest of the paper is organized as follows: We develop the theory of maximizing the uniformity in Sec. 2. We describe the MUD nonlinearity estimation and the entire end-to-end speech recognition system in Sec. 3. Experimental results that demonstrates the effectiveness of the MUD processing is presented in Sec. 4. We conclude in Sec. 5.

Thanks to Samsung Electronics for funding this research. The authors are thankful to Executive Vice President Seunghwan Cho and speech processing Lab. members at Samsung Research.

2. Maximization of Distribution Uniformity

2.1. Power-function based maximization of distribution uniformity

Consider a random variable $\mathbf{X}$ whose range is a closed interval $I_{\mathbf{X}} = [x_{\min}, x_{\max}]$. $x_{\min}$ and $x_{\max}$ are the minimum and maximum values of the random variable $\mathbf{X}$ respectively.

Our objective is to apply a nonlinearity $\sigma_p(\cdot)$ in the form of (1) to X so that the transformed random variable Y closely follows a uniform distribution:

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

We chose the power function as the nonlinearity, partly because it has been shown that this function is quite effective as a compressive nonlinearity in speech feature processing [15, 16, 17, 18]. We subtract X by $x_{\min}$, since this will simplify the maximum likelihood estimation of α, which will be explained shortly. From (1), the range of Y is given by $I_{\mathbf{Y}} = [0, (x_{\max} - x_{\min})^{\alpha}]$. Thus, we expect Y to follow the following uniform distribution:

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

The PDF of Y is given by:

$$p_{\mathbf{Y}}(y) = \begin{cases} \frac{1}{(x_{\max} - x_{\min})^{\alpha}}, & 0 \le y \le (x_{\max} - x_{\min})^{\alpha} \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

Using the property of the PDFs of the transformed random variables [20], we obtain the PDF of the random variable $\mathbf{X}$ by:

$$\begin{aligned} p_{\mathbf{X}}(x) &= p_{\mathbf{Y}}(y) \frac{dy}{dx} \\ &= p_{\mathbf{Y}}(y) \left[\alpha \left(x - x_{\min}\right)^{\alpha - 1} \right] \\ &= \begin{cases} \frac{\alpha (x - x_{\min})^{\alpha-1}}{(x_{\max} - x_{\min})^{\alpha}}, & x_{\min} \le x \le x_{\max} \\ 0, & \text{otherwise.} \end{cases} \end{aligned} \tag{4}$$

Now, suppose that we have the following N samples from the random variable $\mathbf{X}$:

$$X = \{x_0, x_1, \cdots, x_{N-1}\}. \tag{5}$$

Using (4), we obtain the α value which maximizes the data likelihood $p(X|\alpha)$. The log likelihood of the data X assuming the PDF in (2) is given by:

$$\begin{aligned} \mathcal{L}(\alpha|X) &= \sum_{i=0}^{N-1} \ln p_{\mathbf{X}}(x_i) \\ &= \sum_{i=0}^{N-1} \ln \left[\frac{\alpha (x_i - x_{\min})^{\alpha-1}}{(x_{\max} - x_{\min})^{\alpha}} \right] \\ &= N \ln(\alpha) + (\alpha - 1) \sum_{i=0}^{N-1} \ln(x_i - x_{\min}) \\ &\quad - N\alpha \ln(x_{\max} - x_{\min}). \end{aligned} \tag{6}$$

In (7), the term $\ln(x_i - x_{\min})$ is not defined when $x_i = x_{\min}$. Thus, we apply flooring as shown below:

$$\begin{aligned} \mathcal{L}(\alpha|X) &= N \ln(\alpha) \\ &\quad + (\alpha - 1) \sum_{i=0}^{N-1} \ln(\max\{x_i - x_{\min}, \delta\}) \\ &\quad - N\alpha \ln(x_{\max} - x_{\min}), \end{aligned} \tag{7}$$

where δ is a flooring coefficient. We use $\delta = 10^{-100}$ in our experiments. By differentiating $\mathcal{L}(\alpha|X)$ with respect to α, we obtain $\hat{\alpha}$, which maximizes the likelihood as below:

$$\hat{\alpha} = \frac{1}{\ln(\max\{x_i - x_{\min}, \delta\}) - \frac{1}{N}\sum_{i=0}^{N-1} \ln(x_i - x_{\min})}. \tag{8}$$

2.2. Histogram-based maximization of distribution uniformity

Instead of using the power-function based parametric approach to maximize the uniformity of distribution, we may also consider the non-parametric approach. In this approach, we estimate the Cumulative Distribution Function (CDF) from the samples in (5). This CDF estimation may achieved by sorting the samples x_i in (5) and performing interpolation. The relation between the original random variable $\mathbf{X}$ and the transformed random variable $\mathbf{Y}$ is given by the following equation:

$$\mathbf{Y} = \sigma_{np}(\mathbf{X}) = F_u^{-1}\left(\hat{F}_x(\mathbf{X})\right) \tag{9}$$

where $F_u(\cdot)$ is the CDF of the uniform distribution, $\hat{F}_x^{-1}(\cdot)$

3. End-to-End speech recognition with the maximization of feature distribution uniformity

In this section, we explain how to use the theories we developed in Sec. 2.1 and 2.2 to train an end-to-end speech recognition system. The entire block diagram of the system is shown in Fig. 1. We apply either the power function-based MUD nonlinearity in (1) or the histogram-based MUD nonlinearity to each mel filterbank channel as the first step as depicted in Fig. 1. The mel filterbank energy is defined by the following equation:

$$p[m, l] = \sum_{k=0}^{K/2} |X[m, e^{\omega_k}]|^2 M_l[\omega_k] \tag{10}$$

where $M_l[\omega_k]$ is the triangular mel response for the l-th filterbank channel, m is the frame index, and K is the Fast Fourier Transform (FFT) size. ω_k is the discrete-time frequency defined by $\omega_k = \frac{2\pi k}{K}$. The input feature vector $\boldsymbol{x}[m]$ in Fig. 1 is therefore given by:

$$\boldsymbol{x}[m] = [p[m, 0],\ p[m, 1],\ \cdots,\ p[m, L-1]]. \tag{11}$$

where L is the number of mel filter bank channels. In our experiments, we used the value of $L = 40$. For the power function-based MUD, we use (8) for each mel filterbank channels from the randomly selected 1,000 utterances from the training set. In order not to be affected by the silence portion, we removed non-speech portion using a simple energy-based Voice Activity Detector (VAD). Fig. 2 shows the estimated $\hat{\alpha}$ using (8) for each mel filterbank channel. From this Fig. 2, we observe that $\hat{\alpha}$ values are surprisingly close to the power coefficient of $\frac{1}{15}$ which we obtained by modeling the rate-intensity curve using a human auditory system [14, 15]. For the histogram-based MUD, we also used the same randomly selected 1,000 utterances from the training set, applied a VAD, and constructed the empirical CDF to obtain the nonlinearity function in (9). Fig. 3 shows the Probability Density Functions (PDFs) of the mel filterbank energy in (a), those of the nonlinearity output using the power-based MUD in (b), and those of the nonlinearity output

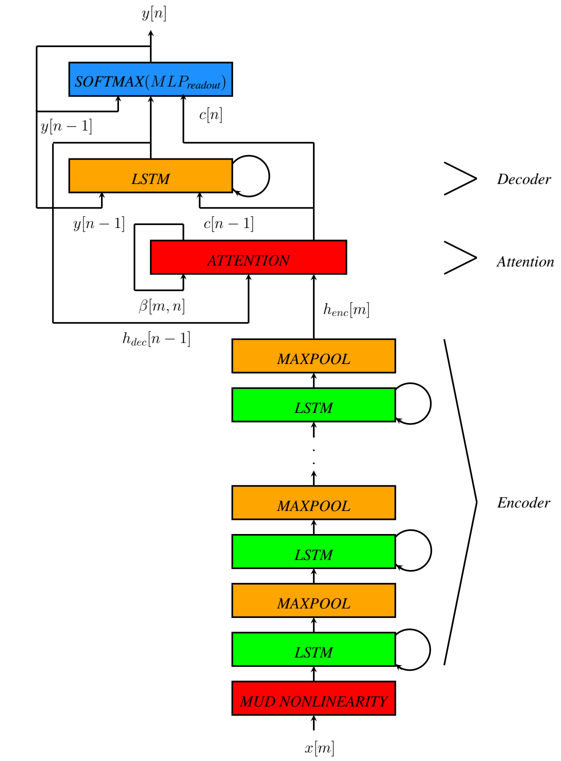


Figure 1: *The structure of the entire end-to-end speech recognition system with MUD processing. The LSTMs in the encoder layers may be either bidirectional-LSTMs or unidirectional-LSTMs. The attention may be either the full attention or the MOnotonic CHunkwise Attention (MOCHA) [21].*

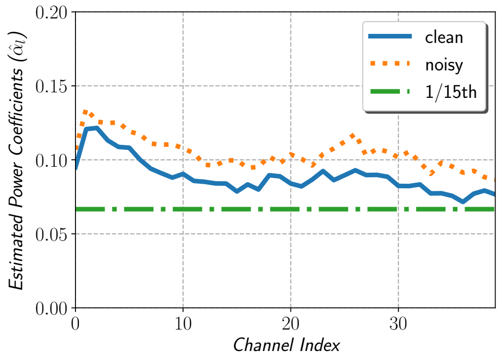
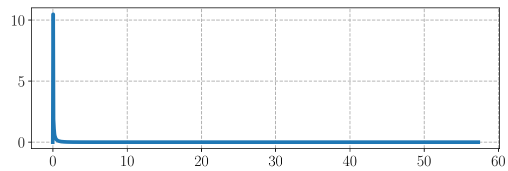
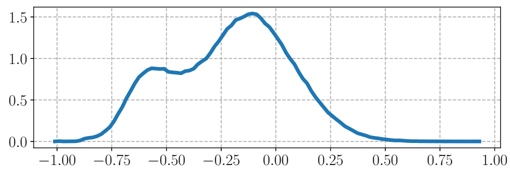


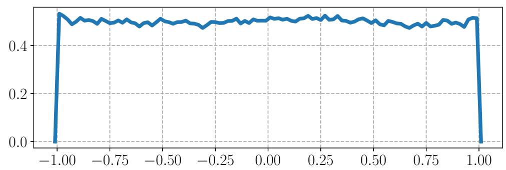
Figure 2: *The estimated power coefficients for each mel filterbank channels using* (8).



(a)



(b)



(c)

Figure 3: *The Probability Density Functions for the third filterbank channel of (a): the mel filterbank energy* $p[m, l], l = 3$ *in* (10)*, (b): the power-function based MUD output of this mel filterbank energy in* (1)*, (c) the histogram based-MUD in* (9)*.*

using the histogram-based MUD in (c). In plotting these PDFs in Fig. 3, we used another 1,000 utterances which are not included in estimating the MUD nonlinearities. These plots are for the third filterbank channel $l = 3$ in (10). As shown in Fig. 3b, if we use the power function-based MUD, the PDF becomes much smoother compared to the original PDF in Fig. 3a. However, this PDF is not as uniform as the one in Fig. 3c. In Fig. 4, we compared the Power-law nonlinearity of the form of $(\cdot)^{\frac{1}{15}}$ used in PNCC [15, 16], power function-based MUD in (1), and histogram-based MUD (9) for the third mel filterbank channel $l = 3$ in (10). Note that in case of power function-based MUD and histogram-based MUD, the nonlinearity is different for different filterbank channels.

We used the RETURNN speech recognition system [23, 24] with various modifications. $\boldsymbol{x}[n]$ $\boldsymbol{y}[n]$ are the input mel filterbank energy vector and the output label , respectively. m is the input frame index and n is the decoder output step index. $\boldsymbol{c}[n]$ is the attention context vector calculated by applying softmax to the attention weights [24]. $\boldsymbol{h}_{enc}[m]$ and $\boldsymbol{h}_{dec}[n]$ are the encoder and the decoder hidden state vectors, respectively. $\boldsymbol{\beta}[m, n]$ is the attention weight feedback [24]. In [24], the peak value of the speech waveform is normalized to be one. However, since finding the peak sample value is not possible for on-line feature extraction, we did not perform this normalization. We modified the input pipeline so that the on-line feature generation can be performed. We disabled the clipping of feature range between -3 and 3, which is the default setting in their Librispeech experiment in [24]. We conducted experiments using both the uni-directional and bi-directional Long Short-Term Memories (LSTMs) [25]. For on-line processing, we used the MOnotonic CHunkwise Attention (MOCHA) [21]. In online speech recognition experiments using MOCHA, we used the chunk size of 2. For better stability in LSTM training, we used the gradient clipping by global norm [26], which is implemented as `tf.clip_by_global_norm` API in Tensorflow [27]. We used six layers of encoders and one layer of decoder followed by a softmax layer.

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 [22]. For each WER number, the same experiment was conducted twice and the results were averaged. All these results were obtained **without** using a Language Model (LM).*

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 %

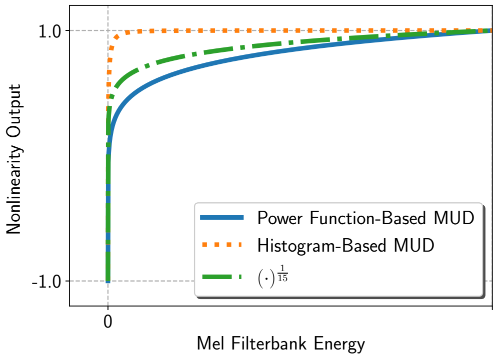


Figure 4: *Comparison of different nonlinearities: Power-law nonlinearity of the form of $(\cdot)^{\frac{1}{15}}$ used in PNCC [15, 16], power function-based MUD in* (1)*, and histogram-based MUD* (9) *for the third mel filterbank channel.*

4. Experimental Results

For speech recognition experiments, we used the Librispeech database [22] for training and evaluation. For training, we used the entire 960 hours training set consisting of 281,241 utterances. For evaluation, we used the official 5.4 hours `test-clean` and 5.1 hours `test-other` databases. We conducted experiments using the 40-th order MFCC feature implemented in [28], power-law nonlinearity of $(\cdot)^{\frac{1}{15}}$ applied to the mel filterbank energy, power function based MUD processing, and histogram-based MUD processing. We conducted experiments using both the online ULSTM/MOCHA [21] structure and the BLSTM with the full-attention structure. We have conducted Bidirectional Long Short-Term Memory (BLSTM) experiments with the cell size of 1536. For the on-line MOCHA experiment, we used the Uni-directional Long Short-term Memory (ULSTM) with the cell size of 1024. In all of our experiments, we did not use any external Language Models (LMs). These results are summarized in Table 1. For each WER number in this table, the same experiment was conducted twice and these results were averaged. The best performance was achieved when we used the power function-based MUD with the 1536-cell BLSTM layers in the encoder and the full attention. For the `test-clean` and test `test-other` test sets [22] , we obtained 4.02 % Word Error Rate (WER) and 13.34 % WER, respectively. On average, the WER was 8.68 %, which is relatively 3.77 % improvement over the baseline MFCC with 9.02 % WER. From Table. 1, we note that usually there is no improvement over the baseline MFCC on the `test-clean` set. However, improvement on the `test-other` was usually more substantial. For the 1536-cell BLSTM full-attention case, the relative improvement over the baseline MFCC on the `test-other` is 4.51 %. The performance difference between the power-law nonlinearity of $(\cdot)^{\frac{1}{15}}$ and the power function-based MUD is usually very small. This was expected since the estimated parameters using (8) are not very different from $\frac{1}{15}$ as shown in Fig. 2. However, for the `test-other` database, which is more a difficult set, the improvement over the power-law nonlinearity of $(\cdot)^{\frac{1}{15}}$ is statistically significant. Histogram-based MUD shows somewhat worse performance compared to power function-based MUD. However, this histogram- based MUD still shows comparable results to the conventional MFCC processing. Unlike the reports in [21], in Table 1, the on-line performance using the MOCHA processing shows significantly worse result than the full attention result. This means that more further optimization of our system is needed for better on-line processing.

5. Conclusions

In this paper, we described the Maximum Uniformity of Distribution (MUD) algorithm. 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. We proposed two different types of MUD approaches: power function-based MUD and histogram -based MUD. In these approaches, we first obtain the Mel filterbank coefficients. The estimated parameters using the power function-based MUD using (8) are surprisingly close to the power coefficient of $\frac{1}{15}$ which we obtained by modeling the rate-intensity curve using a human auditory system [14, 15]. The histogram-based MUD shows comparable performance to the conventional MFCC processing, but it was worse than the performacne of the power function-based MUD. In the end-to-end speech recognition experiments on the Librispeech databases [22], we obtained 4.02 % WER and 13.34 % WER on the `test-clean` and test `test-other` test sets respectively using the power function-based MUD processing. To the best of our knowledge, these are the best results on these databases without using any LMs.

6. References

[1] M. Seltzer, D. Yu, and Y.-Q. Wang, "An investigation of deep neural networks for noise robust speech recognition," in *Int. Conf. Acoust. Speech, and Signal Processing*, 2013, pp. 7398–7402.

[2] D. Yu, M. L. Seltzer, J. Li, J.-T. Huang, and F. Seide, "Feature learning in deep neural networks - studies on speech recognition tasks," in *Proceedings of the International Conference on Learning Representations*, 2013.

[3] V. Vanhoucke, A. Senior, and M. Z. Mao, "Improving the speed of neural networks on CPUs," in *Deep Learning and Unsupervised Feature Learning NIPS Workshop*, 2011.

[4] G. Hinton, L. Deng, D. Yu, 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: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, Nov. 2012.

[5] 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.

[6] 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.

[7] 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, "Raw Multichannel Processing Using Deep Neural Networks," in *New Era for Robust Speech Recognition: Exploiting Deep Learning*, S. Watanabe, M. Delcroix, F. Metze, and J. R. Hershey, Ed. Springer, Oct. 2017.

[8] ——, "Multichannel signal processing with deep neural networks for automatic speech recognition," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, Feb. 2017.

[9] P. Mermelstein, "Automatic segmentation of speech into syllabic units," *J. Acoust. Soc. of Am.*, vol. 58, pp. 880–883, Jan. 1975.

[10] D. Amodei, S. Ananthanarayanan, R. Anubhai, J. Bai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, Q. Cheng, G. Chen *et al.*, "Deep speech 2 : End-to-end speech recognition in english and mandarin," in *Proceedings of The 33rd International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, M. F. Balcan and K. Q. Weinberger, Eds., vol. 48. New York, New York, USA: PMLR, 20–22 Jun 2016, pp. 173–182. [Online]. Available: http://proceedings.mlr.press/v48/amodei16.html

[11] H. Hadian, H. Sameti, D. Povey, and S. Khudanpur, "End-to-end speech recognition using lattice-free mmi," in *Proc. Interspeech 2018*, 2018, pp. 12–16. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2018-1423

[12] C. Kim, E. Variani, A. Narayanan, and M. Bacchiani, "Efficient implementation of the room simulator for training deep neural network acoustic models," 2018, pp. 3028–3032. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2018-2566

[13] T. Sainath, R. Weiss, K. Wilson, A. Narayanan, and M. Bacchiani, "Learning the Speech Front-end With Raw Waveform CLDNNs," in *INTERSPEECH-2015*, Sept. 2015, pp. 1–5.

[14] C. Kim, "Signal processing for robust speech recognition motivated by auditory processing," Ph.D. dissertation, Carnegie Mellon University, Pittsburgh, PA USA, Dec. 2010.

[15] 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.

[16] C. Kim and R. M. Stern, "Power-normalized cepstral coefficients (pncc) for robust speech recognition," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, March 2012, pp. 4101–4104.

[17] ——, "Feature extraction for robust speech recognition based on maximizing the sharpness of the power distribution and on power flooring," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, March 2010, pp. 4574–4577.

[18] ——, "Feature extraction for robust speech recognition using a power-law nonlinearity and power-bias subtraction," in *INTERSPEECH-2009*, Sept. 2009, pp. 28–31.

[19] ——, "Power function-based power distribution normalization algorithm for robust speech recognition," in *IEEE Automatic Speech Recognition and Understanding Workshop*, Dec. 2009, pp. 188–193.

[20] A Papoulis and S. U. Pillai, *Probability, Random Variables and Stochastic Processes*. New York, NY: McGraw-Hill, 2002.

[21] C.-C. Chiu and C. Raffel, "Monotonic chunkwise attention," in *International Conference on Learning Representations*, Apr. 2018. [Online]. Available: https://openreview.net/forum?id=Hko85plCW

[22] 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.

[23] P. Doetsch, A. Zeyer, P. Voigtlaender, I. Kulikov, R. Schlüter, and H. Ney, "RETURNN: the RWTH extensible training framework for universal recurrent neural networks," in *IEEE Int. Conf. Acoust., Speech, and Signal Processing*, March 2017, pp. 5345–5349.

[24] 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

[25] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, no. 9, pp. 1735–1780, Nov. 1997.

[26] R. Pascanu, T. Mikolov, and Y. Bengio, "On the difficulty of training recurrent neural networks," in *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28*, ser. ICML'13. JMLR.org, 2013, pp. III–1310–III–1318. [Online]. Available: http://dl.acm.org/citation.cfm?id=3042817.3043083

[27] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, M. Kudlur, J. Levenberg, R. Monga, S. Moore, D. G. Murray, B. Steiner, P. Tucker, V. Vasudevan, P. Warden, M. Wicke, Y. Yu, and X. Zheng, "Tensorflow: A system for large-scale machine learning," in *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*. Savannah, GA: USENIX Association, 2016, pp. 265–283. [Online]. Available: https://www.usenix.org/conference/osdi16/technical-sessions/presentation/abadi

[28] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," in *Proceedings of the 14th Python in Science Conference*, K. Huff and J. Bergstra, Eds., 2015, pp. 18 – 25.