NOISY OBJECTIVE FUNCTIONS BASED ON THE F-DIVERGENCE

Markus Nussbaum-Thom[1,2]*, Ralf Schlüter*[2]*, Vaibhava Goel*[1]*, Hermann Ney*[2,3]

[1] IBM T.J. Watson Research Center, Yorktown Heights, NY, 10598
[2] Computer Science Dept. 6, RWTH Aachen University, Aachen, Germany
[3] Spoken Language Processing Group, LIMSI CNRS, Paris, France
{nussbaum, vgoel}@us.ibm.com, {schlueter, ney}@i6.informatik.rwth-aachen.de

ABSTRACT

Dropout, the random dropping out of activations according to a specified rate, is a very simple but effective method to avoid over-fitting of deep neural networks to the training data.

In this work, we approach regularization from the view of the objective function by dynamically changing the objective function to avoid local minima. The underlying theory is based on our previous work where we showed that a novel family of training criteria exists based on the f-divergence. These criteria are a generalization of the cross-entropy criterion. We introduce two regularization schemes – the first approach minimizes over a family of training criteria in order to achieve the lowest possible criterion, and the second approach randomly chooses a criterion from a family of criteria according to a Gaussian distribution.

In practical experiments on the WSJ-5K corpus, the proposed schemes are successfully evaluated compared to dropout for deep neural networks and bidirectional gated recurrent units, both as standalone approaches and in combination with dropout.

Index Terms— dropout, generalization error bounds, training criteria, f-divergence, neural networks

1. INTRODUCTION

It is well known that the training of deep neural networks suffers from over-fitting to the training data. Several techniques such as weight regularization (L1 and L2), weight constraints, early stopping and model averaging can make the over-fitting to the training data less damaging. Dropout, the random dropping out of activation output, has become the most popular among these techniques [1].

In this work, two new regularization techniques are proposed. The novel techniques dynamically change the training criterion to avoid getting stuck in local minima too early. This approach is based on our previous work in [2, 3, 4] where classification error bounds and related training criteria based on the *f-Divergence* are derived.

In [4], the Conjugate Power Approximation criterion (α-CPA) with parameter $\alpha \in [0, 1]$ was derived from classification error bounds based on the *f-Divergence*. This class of criteria is a special case of the cross-entropy (CE) criterion i.e. the power approximation converges to the logarithm for $\alpha \to 0$. Practical experiments suggested that a combination of the conjugate power approximation and the cross-entropy criterion can result in an improved criterion by tuning over α.

Here, we propose two scheduling approaches for choosing α –

- **Minimum Conjugate Power Approximation**: Minimize α-CPA over α for each sample or mini-batch.
- **Noisy Objective Function**: Randomly draw α according to a Gaussian distribution with a specific mean (usually close to zero) and a specific variance for each sample or mini-batch.

In practical experiments, we compare and combine both schemes with the state-of-the-art dropout regularization technique. We train and evaluate Deep Neural Networks (DNNs) and Bidirectional Gated Recurrent Units (BGRUs) as acoustic models using different regularization techniques on the WSJJ0 speech recognition task. The proposed schemes are successfully compared to dropout, both as standalone approaches and in combination with dropout.

Section 2 reviews our previous work from [2, 3, 4]. Sections 3 and 4 introduce the novel minimum conjugate power approximation and noisy objective function approach. Sections 5 and 6 discuss the experimental setup and results The paper concludes with Section 7.

2. A FAMILY OF DISCRIMINATIVE TRAINING CRITERIA BASED ON THE F-DIVERGENCE

In this section our previous work from [2, 3, 4] is briefly reviewed. Consider a statistical classification problem with a model distribution $q(x, c)$ of the continuous observations $x \in \mathcal{X}$ and classes $c \in \mathcal{C}$, which is used to classify samples of the unknown true distribution $pr(x, c)$. The *Bayes* $c_{pr}(x)$ and model $c_q(x)$ decision rules corresponding to the true and model posterior probabilities $pr(c|x)$ and $q(c|x)$ are defined as

$$c_{pr}(x) := \operatorname*{argmax}_{c \in \mathcal{C}} \{pr(c|x)\}$$
$$c_q(x) := \operatorname*{argmax}_{c \in \mathcal{C}} \{q(c|x)\}.$$

The quality of the model is measured by the local and global classification error difference associated with the decision rules

$$\Delta(x) := pr(c_{pr}(x)|x) - pr(c_q(x)|x)$$
$$\Delta := \int pr(x)\Delta(x)\ \mathrm{d}x.$$

Definition 1 *If $f : \mathbb{R}^+ \to \mathbb{R}$ is a convex and $f(1) = 0$ then*

$$D_f^x(pr||q) := \sum_{c \in \mathcal{C}} q(c|x) f\left(\frac{pr(c|x)}{q(c|x)}\right)$$

is defined as the f-Divergence [5, 6, 7].

In [2, 3], the *f-Divergence* was introduced to derive the following tight implicit classification error bound on the global and local classification error difference

$$2D_f^x(pr||q) \geq f(1 + \Delta(x)) + f(1 - \Delta(x)). \quad (1)$$

There, this local implicit classification error bound was extended to explicit global bounds for a specific type of *f-Divergence* functions. If $f : \mathbb{R}^+ \to \mathbb{R}$ is convex, $f(1) = 0$, and $f'''(u)$ exists and is monotonically increasing in $u \in [0, 1]$, then the following expression establishes a bound on the global classification error difference

$$\Delta^2 \leq \frac{1}{f''(1)} \int pr(x) D_f^x(pr||q) \, \mathrm{d}x.$$

For conjugate convex functions $f(u) = u \cdot g(1/u)$, and if $g(u)$ is monotonically decreasing, this bound can be transformed into an empirical training criterion $F_f(q)$ using the empirical distribution on the labeled samples $(x_n, c_n), n = 1, \ldots, N$

$$2\frac{1}{f''(1)} \int pr(x) D_f^x(pr||q) \, \mathrm{d}x$$

$$\rightsquigarrow F_f(q) = \frac{1}{f''(1)} \frac{1}{N} \sum_{n=1}^{N} g(q(c_n|x_n)).$$

The conjugate power approximation (α-CPA) training criterion associated with the function $f(u) = -u\left(\frac{1}{u^\alpha} - 1\right)$ and $f''(1) = 1 - \alpha$ fulfills the above conditions and results in

$$\mathcal{F}_{\alpha\text{-CPA}}(q) = \frac{1}{N} \sum_{n=1}^{N} \frac{(1 - q^\alpha(c_n|x_n))}{\alpha(1-\alpha)}. \tag{2}$$

For $\alpha \to 0$ the power approximation converges to the logarithm $\log(u) = \frac{u^\alpha - 1}{\alpha}$, while the corresponding *f-Divergence* converges to the *Kullback-Leibler f-Divergence*, and the criterion converges to the cross-entropy criterion.

$$\mathcal{F}_{\text{CE}}(q) = -\frac{1}{N} \sum_{n=1}^{N} \log q(c_n|x_n) \tag{3}$$

As convex functions are closed under addition, this is also valid for the corresponding *f-Divergences* and derived training criteria.

In the next section we introduce the minimum conjugate power approximation training criterion which minimizes over α to derive a smaller bound, and therefore a better training criterion.

3. MINIMUM CONJUGATE POWER APPROXIMATION

All training criteria are derived from a classification error bound on the classification error difference. By choosing a tighter bound we can expect a better training criterion. According to this argument, Figure 1 shows that for different model posteriors different α-CPA criteria are minimal. Therefore, by minimizing the corresponding α-CPA classification error bound over α will result in a tighter bound. The corresponding training criteria for a sample or batch-wise minimization over α results in

$$\frac{1}{N} \sum_{n=1}^{N} \min_{\alpha \in [0,\beta]} \left\{ \frac{(1 - q^\alpha(c_n|x_n))}{\alpha(1-\alpha)} \right\} \qquad \text{(MIN-SAMP-CPA)}$$

$$\min_{\alpha \in [0,\beta]} \left\{ \frac{1}{N} \sum_{n=1}^{N} \frac{(1 - q^\alpha(c_n|x_n))}{\alpha(1-\alpha)} \right\} \qquad \text{(MIN-BATCH-CPA)}$$

For practical purposes we have included a parameter $\beta \in [0, 1]$ to limit the choice of α. The practical implementation uses a golden section search for the minimal α within the interval $[0, \beta]$.

The next section introduces the other proposed approach of noisy objective functions which chooses the parameter α randomly drawn according to a Gaussian distribution.

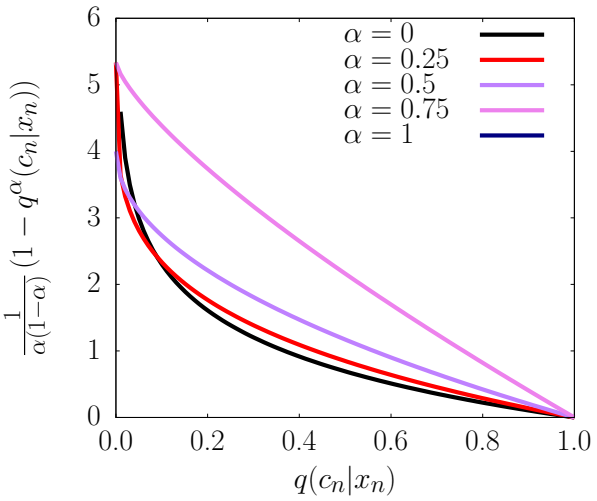

Fig. 1. *Function* $\frac{1-q(c_n|x_n)}{\alpha(1-\alpha)}$ *corresponding to one sample of the* α*-CPA criterion which converges to the cross-entropy criterion for* $\alpha \to 0$.

4. NOISY OBJECTIVE FUNCTION

In this section we introduce the noisy conjugate power approximation criterion. By randomly choosing the parameter $\alpha \sim \mathcal{N}(\mu, \sigma^2)$ according to a Gaussian distribution with specific mean μ and variance σ^2, the resulting training criterion can be expected to be less sensitive to local minima. Therefore, this method is suitable for regularization towards a better local optimum. The resulting training criterion randomly chooses α either per sample or batch-wise

$$\frac{1}{N} \sum_{n=1}^{N} \underset{\substack{\alpha \sim \mathcal{N}(\mu,\sigma^2) \\ \alpha \in [0,1]}}{\text{rand}} \left\{ \frac{(1 - q^\alpha(c_n|x_n))}{\alpha(1-\alpha)} \right\} \qquad \text{(RAND-SAMP-CPA)}$$

$$\underset{\substack{\alpha \sim \mathcal{N}(\mu,\sigma^2) \\ \alpha \in [0,1]}}{\text{rand}} \left\{ \frac{1}{N} \sum_{n=1}^{N} \frac{(1 - q^\alpha(c_n|x_n))}{\alpha(1-\alpha)} \right\} \qquad \text{(RAND-BATCH-CPA)}$$

5. EXPERIMENTAL SETUP

The Gaussian mixture Hidden Markov Model (GHMM) baseline recognition system for WSJ0 uses 1500 generalized triphone states which were top down clustered using a decision tree, plus one silence state. The corpus statistics for WSJ0 are shown in Table 1. The emission probabilities are modeled by Gaussian mixture distributions with a total of about 200k densities. The raw acoustic features are 19-dimensional PLP features. Temporal context is included by splicing 9 successive frames of PLP features into super-vectors,

Table 1. Corpus statistics (RW : running words).

Corpus	Train/ Dev/ Eval		
	Data[h]	#Segments	#Words
WSJ0	15:17/ 0:46/ 0:4	7k/ 410/ 330	130k/ 6k/ 5k

then projecting to 40 dimensions using linear discriminant analysis (LDA). For recognition purpose a 5k lexicon and trigram language model for WSJ0 is used.

All neural network experiments use an acoustic front-end that comprises of 40 Log-Mel features augmented with delta and double delta, and are evaluated as hybrid acoustic models for automatic speech recognition. Deep Neural Networks (DNNs) and Bidirectional Gated Recurrent Units (BGRUs) are trained using the architectures and recipes described in [8].

6. EXPERIMENTAL RESULTS

In this section the result of a series of experiments which combine and compare different training criteria are described. In the following we use the notation A+B in case a criterion A is combined with a criterion B, i.e. A+B = (A+B)/2 is effectively used for training. The training criteria derived here are still derived from error bounds based on the *f-Divergence*, as *f-Divergences* are closed under addition. In the subsequent experiments the best model is always chosen according to best the WER on the DEV corpus. As a contrastive result, the overall best model on the EVAL corpus achieves a WER of 1.9%. This is a BGRU model with dropout 0.1 and is trained using the CE+RAND-SAMP-CPA criterion. This model however achieves a WER of 2.6% on the DEV corpus.

Table 2. The WER[%] as a function of different training criteria working as a regularization scheme for DNN, BGRU and BGRU(0.1) (BGRU with dropout rate 0.1) model on the WSJ0 test corpora.

MODEL	CRITERION	WER[%]	
		DEV	EVAL
DNN	CE	3.1	3.3
	MIN-SAMP-CPA	3.0	3.5
	CE+MIN-SAMP-CPA	**2.9**	**3.0**
	MIN-BATCH-CPA	3.1	3.4
	CE+MIN-BATCH-CPA	3.1	3.2
	RAND-SAMP-CPA	3.0	3.3
	CE+RAND-SAMP-CPA	**2.9**	**3.0**
	RAND-BATCH-CPA	2.9	3.5
	CE+RAND-BATCH-CPA	**2.9**	**3.1**
BGRU	CE	2.6	2.4
	MIN-SAMP-CPA	2.7	2.0
	CE+MIN-SAMP-CPA	2.7	2.2
	MIN-BATCH-CPA	2.7	2.2
	CE+MIN-BATCH-CPA	2.7	2.1
	RAND-SAMP-CPA	2.6	2.2
	CE+RAND-SAMP-CPA	**2.6**	**2.0**
	RAND-BATCH-CPA	2.5	2.2
	CE+RAND-BATCH-CPA	**2.5**	**2.2**
BGRU(0.1)	CE	2.6	2.3
	MIN-SAMP-CPA	2.6	2.3
	CE+MIN-SAMP-CPA	2.7	2.1
	MIN-BATCH-CPA	2.6	2.2
	CE+MIN-BATCH-CPA	2.7	2.2
	RAND-SAMP-CPA	2.6	2.1
	CE+RAND-SAMP-CPA	**2.6**	**2.1**
	RAND-BATCH-CPA	2.5	2.3
	CE+RAND-BATCH-CPA	**2.5**	**2.2**

In initial experiments, the DNNs are first trained using the CE, CPA and CE+α-CPA for $\alpha \in \{0.001, 0.01, 0.1, 0.2, 0.3, 0.4\}$. The baseline CE criterion using the DNN model results in a WER of 3.1% on the DEV corpus, and 3.3% on the EVAL corpus.

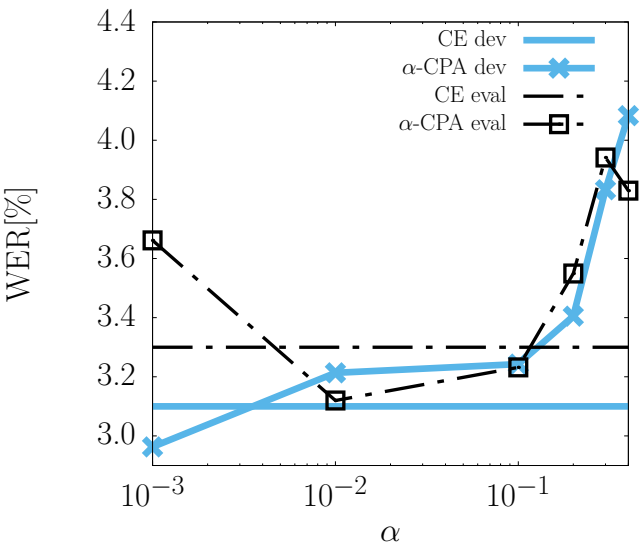

Fig. 2. *The WER[%] as a function of the parameter α of the α-CPA criterion trained using DNN models.*

Figure 2 shows the result for the α-CPA criterion compared to the CE criterion only. The CE+α-CPA criterion is not shown since as it had a similar performance.

By choosing the optimal α according to the WER on the DEV corpus, a small improvement over the CE criterion is achieved while leading to a WER degradation on the EVAL corpus. For slightly different $\alpha \in \{0.01, 0.1\}$, the EVAL performance is better than the baseline CE criterion. The evaluation of the MIN-CPA and CE+MIN-CPA criterion minimized on a sample or batch-wise, without restriction on β (i.e. $\beta = 1$), results in a WER ranging from 3.7 to 4.0. This is slightly worse than the CE baseline criterion. The analysis of these results, and with Figure 1 in mind, where the CE criterion is lower-bounded by α-CPA for smaller values of α, suggests a different strategy – a smaller α close to zero can achieve a better criterion. Therefore, in the next set of experiments, α is constrained to smaller values either by choosing a small β, or choosing a smaller mean and variance.

In the next set of experiments we apply dropout to the intermediate layer outputs for DNNs and BGRUs using the dropout rates $\{0, 0.01, 0.1, 0.2, 0.3, 0.4, 0.5\}$. For BGRUs we also found it useful to apply the same dropout rate to the weights simultaneously. Dropout did not help for DNNs. However, for BGRUs the best dropout rate is 0.1 which reduced the WER on the DEV corpus only slightly by a couple of error counts, but reduced the WER on the EVAL corpus from 2.4 to 2.3.

In the next set of experiments we train DNN and BGRU models with a variety of combined criteria where the choice for α is constrained to values close to zero. For $\beta \in \{10^{-i} | i \in \{1, 2, 3, 4, 5\}\}$ the DNN and BGRU models are trained using the following criteria

- MIN-SAMP-CPA,
- CE+MIN-SAMP-CPA,
- MIN-BATCH-CPA,

- and CE+MIN-BATCH-CPA.

Also for $\mu \in \{10^{-i} | i \in \{1,2,3,4,5,6\}\}$ and $\sigma^2 \in \{10^{-i} | i \in \{1,2,3,4,5,6\}\}$ the DNN and BGRU models are trained using the following criteria

- RAND-SAMP-CPA,
- CE+RAND-SAMP-CPA,
- RAND-BATCH-CPA,
- and CE+RAND-BATCH-CPA.

Figures 3 and 4 show the WER curve of the CE+MIN-SAMP-CPA criterion and the CE+RAND-SAMP-CPA criterion for BGRU models with dropout 0.1 respectively.

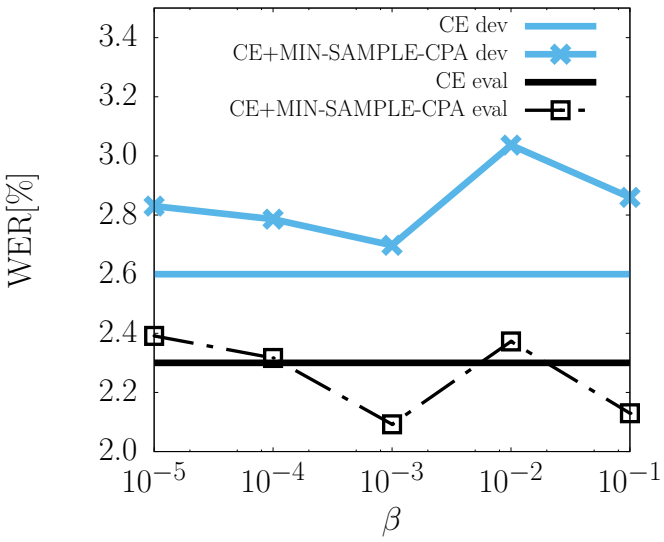

Fig. 3. *The WER[%] as a function of the constraint β for the CE+MIN-SAMP-CPA criterion for BGRU models with dropout 0.1 on the* WSJ0 *test corpora.*

Table 2 shows the result of different training criteria combinations for the various models. The best results have been highlighted. In summary, using the combined CE+MIN-SAMP-CPA, CE+MIN-BATCH-CPA, CE+RAND-SAMP-CPA, CE+RAND-BATCH-CPA criteria for both DNN and BGRU models, results in a similar or lower WER on the DEV corpus, and a 0.2-0.4% WER improvement on the EVAL corpus. This is a 9-20% relative improvement on the EVAL corpus. For BGRU models the influence of dropout in combination with the novel criteria has no major impact on the WER. However, a small negative impact on the WER is observed in cases where the combination does not include the CE criterion. Overall, the randomized criteria CE+RAND-SAMP-CPA and CE+RAND-BATCH-CPA perform more stably than the other criteria.

7. CONCLUSION

Two novel regularization techniques were introduced based on new training criteria. Following a principled approach, training criteria are derived from *f-Divergence* bounds on the classification error difference between the *Bayes* and model decision rule. The first technique minimizes over this family of training criteria. This continuously changes the functional form of the training criterion to avoid

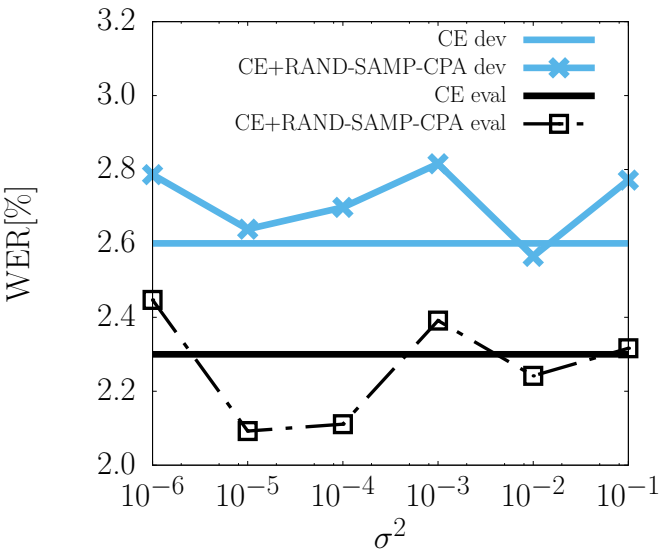

Fig. 4. *The WER[%] as a function of the variance σ^2 for the CE+RAND-SAMP-CPA criterion with $\mu = 10^{-6}$ for BGRU models with dropout 0.1 on the* WSJ0 *test corpora.*

local minima too early. The second technique dynamically changes the objective function by randomly drawing from a family of criteria according to a Gaussian distribution which also avoids local minima. Both techniques successfully showed a WER improvement over the cross entropy baseline criterion when tested in practical experiments on the WSJ-5K corpus for deep neural networks and bidirectional gated recurrent units, both in standalone implementations and in combination with dropout.

8. ACKNOWLEDGMENT

The research leading to these results has received funding from the Intelligence Advanced Research Projects Activity (IARPA) via Department of Defense U.S. Army Research Labotory (DoD/ARL) contract no. W911NF-12-C-0012. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes not withstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DOD/ARL, or the U.S. Government.

9. REFERENCES

[1] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov, "Dropout: A Simple Way to Prevent Neural Networks from Overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, Jan. 2014.

[2] M. Nußbaum-Thom, E. Beck, T. Alkhouli, R. Schlüter, and H. Ney, "Relative Error Bounds for Statistical Classifiers Based on the f-Divergence," in *Interspeech*, Lyon, France, Aug. 2013.

[3] Ralf Schlüter, Markus Nussbaum-Thom, Eugen Beck, Tamer Alkhouli, and Hermann Ney, "Novel Tight Classification Error Bounds under Mismatch Conditions based on f-Divergence," in

Information Theory Workshop (ITW), 2013 IEEE. IEEE, 2013, pp. 1–5.

[4] Markus Nußbaum-Thom, Xiaodong Cui, Ralf Schlüter, Vaibhava Goel, and Hermann Ney, "A Family of Discriminative Training Criteria based on the f-Divergence for deep neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2014, Florence, Italy, May 4-9, 2014*, 2014, pp. 5612–5616.

[5] I. Csiszár and P. C. Shields, "Information Theory and Statistics: A Tutorial," vol. 1, no. 4, 2004.

[6] F. Liese and I. Vajda, "On Divergences and Informations in Statistics and Information Theory," vol. 52, no. 10, pp. 4394–4412, 2006.

[7] F. Österreicher, "Csizár's f-Divergences - Basic Properties," in *Talk presented at workshop of the Research Group in Mathematical Inequalities and Applications at the Victoria University*, Melbourne, Australia, Oct. 2002.

[8] Markus Nussbaum-Thom, Jia Cui, Bhuvana Ramabhadran, and Vaibhava Goel, "Acoustic Modeling using Bidirectional Gated Recurrent Convolutional Units," *Interspeech 2016*, pp. 390–394, 2016.