MULTI-ACCENT SPEECH RECOGNITION WITH HIERARCHICAL GRAPHEME BASED MODELS

Kanishka Rao, Haşim Sak

Speech Group, Google Inc.
{kanishkarao,hasim}@google.com

ABSTRACT

We train grapheme-based acoustic models for speech recognition using a hierarchical recurrent neural network architecture with connectionist temporal classification (CTC) loss. The models learn to align utterances with phonetic transcriptions in a lower layer and graphemic transcriptions in the final layer in a multi-task learning setting. Using the grapheme predictions from a hierarchical model trained on 3 million US English utterances results in 6.7% relative word error rate (WER) increase when compared to using the phoneme-based acoustic model trained on the same data. However, we show that hierarchical grapheme-based models trained on larger acoustic data (12 million utterances) jointly for grapheme and phoneme prediction task outperform phoneme only model by 6.9% relative WER. We train a single multi-dialect model using a combined US, British, Indian and Australian English data set and then adapt the model using US English data only. This adapted multi-accent model outperforms a model exclusively trained on US English. This process is repeated for phoneme-based and grapheme-based acoustic models for all four dialects and larger improvements are obtained with grapheme models. Additionally using a multi-accent grapheme model, we observe large recognition accuracy improvements for Indian-accented utterances in Google VoiceSearch US traffic with a 40% relative WER reduction.

Index Terms: deep neural networks, grapheme, acoustic modelling, CTC

1. INTRODUCTION

The current state-of-the-art automatic speech recognition (ASR) systems consist of three components; acoustic model, pronunciation model and language model. These three components are typically trained on different datasets and optimized independently from the others. For example, the acoustic model is trained on acoustic data to predict phonetic labels for each acoustic feature vector [1] while the language model is trained on text data to predict the next word in a sequence of words [2]. Finally, these three components are combined typically in a finite-state transducer based speech recognition system and accuracy is evaluated using word error rate (WER). Given recent advances in deep learning [3, 4, 5, 6] for speech recognition, an exciting prospect is unifying these models into an end-to-end optimized system for recognition accuracy.

Recently, we have seen dramatic improvements in acoustic modeling with the transition from Gaussian Mixture Models [7] to Deep Neural Networks [8] and more recently to Recurrent Neural Networks (RNN) [9]. Similarly, language models have been improved with RNNs [10, 11, 12]. However, the pronunciation model in state-of-the-art systems still remains relatively simple consisting of a dictionary of human transcribed word pronunciations with a grapheme-to-phoneme [13] model as a back-off. Such a dictionary based model may suffer from several problems:

- Dictionary is not a statistical model; if a word has multiple pronunciations these are simply listed as multiple entries in the dictionary.
- Dictionary is defined at the word level and any inter-word coarticulation effects are not considered.
- Pronunciations are defined for slow speech, real fast speech may skip certain phonemes.
- Pronunciations are defined for a particular accent.

In an end-to-end optimized system, the pronunciation model may be learned along with the acoustic model [14, 15] directly from acoustic data and thus may capture the irregularities of pronunciations for spoken words. A popular approach to accomplish this is with character or grapheme-based acoustic model [16, 17, 18] recognizing graphemes instead of phonemes in the traditional setup. Although grapheme-based recognition has previously been shown to work, it generally performs worse than traditional phoneme-based systems[19, 20]. Grapheme-based recognition has also been applied to multi-lingual recognition [21] since such grapheme based units can be shared across languages.

In this paper, we confirm that phoneme-based recognition combined with an expert-written pronunciation dictionary typically outperforms a grapheme-based recognition with no pronunciation dictionary, both trained on roughly 3000 hours of transcribed speech data. The primary goal of this work is to investigate how grapheme-based recognition may scale with larger data sets and model sizes in order for them to be competitive with the phoneme system with a curated pronunciation dictionary.

To create a grapheme recognizer we use the connectionist temporal classification (CTC) [22] technique to train models to directly predict graphemes. We also train models with hierarchical CTC [23] in a multi-task learning setting and show improvements over non-hierarchical CTC models. To explore how the accuracy of grapheme and phoneme models change with large data sets we consider a single acoustic training data set where we combine various dialects of English from US, UK, India and Australia. As a result, we create an accent-robust ASR system by training a single grapheme-based model on this multi-dialect data set. Finally, we show that when trained with additional data grapheme-based models outperform phoneme-based models.

2. RNN ACOUSTIC MODELING TECHNIQUES

2.1. LSTM

Recurrent neural networks, specifically Long Short-Term Memory (LSTM) networks have proven to be the sate-of-the-art in numerous

sequence modeling tasks [9]. LSTM can model long range temporal dependencies which makes them particularly suited for speech recognition [24]. Bidirectional LSTM [9] acoustic models process the input in both the forward and backward directions and model both the right and left temporal context and perform better than unidirectional models when you can afford the latency of processing the entire speech input before estimating label posteriors. In this work, we exclusively train networks as stacked bidirectional LSTM layers where at each depth two LSTM layers (one forward and one backward) are fully connected to the two LSTM layers at the next adjacent depth [24].

2.2. CTC

The CTC approach [22] is a technique for sequence labeling using RNNs where the alignment between the inputs and target labels is unknown. CTC can be implemented with a softmax output layer using an additional unit for the *blank* label used to estimate the probability of outputting no label at a given time. The output label probabilities from the network define a probability distribution over all possible labelings of input sequences including the *blank* labels. The network can be trained to optimize the total log probability of correct labelings for training data as estimated using the network outputs and forward-backward algorithm [25]. The correct labelings for an input sequence are defined as the set of all possible labelings of the input with the target labels in the correct sequence possibly with repetitions and with *blank* labels permitted between separate labels.

2.3. Grapheme-based Models

We use the CTC objective function to train a single network to directly predict graphemes given the acoustic input. We choose the lower cased English alphabet (a-z) as the grapheme target labels. A special <*space*> label indicates boundary between words making a total of 27 grapheme labels. For the training labels we convert transcriptions to the spoken domain using a verbalizer [26]. This verbalizer is constructed manually based on language specific rules and may generate several alternative spoken transcriptions for a given written transcription, for example, $101 $\rightarrow$ *one hundred and one dollars* or *hundred and one dollars* or *one oh one dollars*. The grapheme sequence in the chosen spoken form that aligns best with the audio using a speech recognizer is used as the CTC grapheme targets, such as *hundred and one dollars* $\rightarrow$ *h u n d r e d* <*space*> *a n d* <*space*> *o n e* <*space*> *d o l l a r s*.

2.4. Hierarchical Grapheme Models

Speech recognition can be formulated as a hierarchical sequence modeling task with the modeling of acoustics to phonemes and modeling of phonemes to graphemes as two hierarchies in a multi-task learning setting. Multi-task acoustic modelling has been used previously to improve phoneme recognition [27, 28] by modeling additional phonetic knowledge as a secondary task. In this work we attempt to improve grapheme recognition by modeling phonemes as a secondary task. Hierarchical models can be trained with CTC [23] with multiple CTC losses estimated at various depths in the network. We create a hierarchical-CTC grapheme recognizer with a phoneme CTC loss in an intermediate layer, see Figure 1. The CTC losses are backprogapated from both output layers and the gradients are added for the layers before phoneme output, this is similar to using a $\lambda = 1$ as described in [23]. Note that with such a hierarchical CTC setup the human-transcribed pronunciation dictionaries are still used to generate targets for the phoneme output, however, this is only used during training and the grapheme outputs are used for inference. We find that using hierarchical-CTC loss which also requires phoneme outputs significantly improves the performance of the grapheme recognizer and this is further discussed in section 4.

2.5. Multi-Dialect Hierarchical Grapheme Models

For robustness to different accents and to increase the amount of acoustic training data we combine English data across 4 locales; US, UK, India and Australia. Similar techniques for training multilingual acoustic models on combined data have been used successfully to increase performance [29, 30], in this work we train a single neural network on this combined data of 4 locales. We create a modified version of the hierarchical-CTC network, see Figure 2, for this multi-dialect data. Four phoneme output layers, one for each locale, are fully connected to an intermediate layer in addition to the final grapheme output layer. CTC loss is only estimated on the phoneme layer matching the locale of the utterance in addition to the grapheme CTC loss. Four different locale-specific human-transcribed pronunciation dictionaries are used to generate the targets for the phoneme outputs. We found that using a single pronunciation dictionary for all the data slightly hurt the model performance.

2.6. Sequence Training

CTC loss is not optimal for WER and it has been shown that sequence-level discriminative training criteria incorporating the language model further improve the performance of RNN acoustic models trained with CTC [22]. In this paper, we use the state-level minimum Bayes risk (sMBR) sequence discriminative training criterion [31] to further improve the accuracy of CTC-trained models. For all hierarchical-CTC models the intermediate loss is discarded and only the final output layer is used for sequence training.

3. EXPERIMENTAL SETUP

We train and evaluate models on hand-transcribed, anonymized utterances taken from Google voice search US English traffic. Our training and test sets consist of 3 million and 28 thousand utterances, respectively, each utterance being roughly 4s. Datasets of the same sizes are also collected for British English, Indian English and Australian English. A special test set of 25,000 anonymized utterances is created from US English traffic consisting of Indian accent utterances. A classifier was used to select such accented utterances which were then validated and transcribed by humans. Evaluation uses a 5-gram language model pruned to 100 million n-grams. Rescoring of word lattices is done with a larger n-gram model pruned to 15 billion n-grams.

We use 80-dimensional log mel filterbank energy features computed every 10ms on 25ms windows. We first stack frames so that the networks sees multiple (8) frames at a time but then decimate the frames so that we skip forward multiple frames (3) after processing each such super-frame. This setup is described in more detail in [24].

We train a grapheme recognizer (grapheme-based acoustic model) with 5 layers of bidirectional LSTM with a CTC loss function. For the hierarchical models we use 8 layers of bidirectional LSTM with the primary CTC loss on the 5th layer and the secondary CTC loss on the final 8th layer. We experimented with architectures of 3 to 10 layers deep with the primary loss at the 3rd to 10th layer and found no futher improvements after 8 layers and the primary

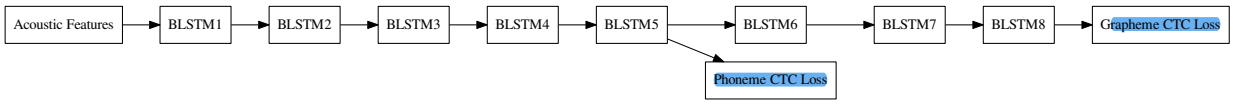

Fig. 1: A hierarchical LSTM-RNN with CTC losses based on graphemes and phonemes.

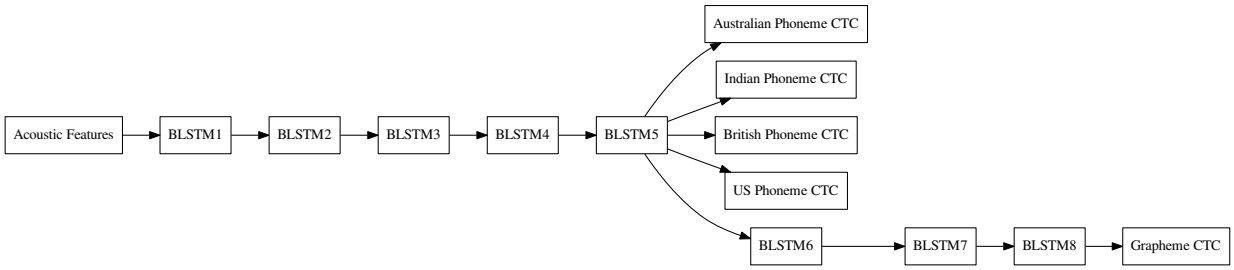

Fig. 2: A hierarchical LSTM-RNN with CTC losses based on graphemes and multi-dialect phonemes.

loss at the 5th layer. Each bidirectional LSTM layer consists of 300 memory cells for each direction. All models are randomly initialized with a uniform weight (-0.04, 0.04) distribution and trained in a distributed manner using asynchronous stochastic gradient descent (ASGD) with a learning rate of 0.0001. To stabilize CTC training we clip the activations of memory cells to [-100, 100], and their gradients to [-1, 1]. A held-out set from the training data is used to determine the early stopping for each model.

For phoneme targets we use the XSampa phoneset with 42 phonemes for US English, for grapheme targets we use the lowercase English alphabet, a-z, along with a $<space>$ label totaling 27 labels.

4. RESULTS AND DISCUSSION

4.1. Single Dialect Models

This section only contains results with models trained on a single dialect training data. We find that a single 5-layer grapheme-CTC network achieving 12.5% WER performs poorly compared to a single 5-layer phoneme-CTC network with 10.5% WER. The 8-layer hierarchical grapheme improves the performance from 12.5% $\rightarrow$ 11.2%. To make sure that this improvement was not due to simply increasing the size of the network we trained an 8-layer hierarchical phoneme model (with phoneme CTC losses at both hierarchies) and found little impact on the WER, 10.5% $\rightarrow$ 10.6%. It seems that the grapheme recognizer benefits from phonetic information provided as phoneme targets. Since CTC-P and HCTC-G were found to be the optimal architectures for phoneme and grapheme models respectively we will only compare those two in the following sections.

For each of the four dialects we train phoneme-based and grapheme based CTC models using the respective 3 million utterances. The WERs are summarized in Table 1. The grapheme based models perform worse than phoneme models for all languages with the largest relative regression being 10.3% for EnGB and smallest 0.7% for EnAu.

4.2. Multi Dialect Models

We combine the four 3-million utterance datasets into a single multi-dialect 12-million utterance English dataset and train a single multi-dialect grapheme model with a modified hierarchical-CTC approach as previously described in Section 2.5. We find this single grapheme model outperforms (see Table 1) the individual locale-specific phoneme models for EnUs, EnGb and EnAu and has a small 1.3% relative regression for EnIn.

Finally, we adapt the single multi-dialect grapheme model for a specific locale by continuing sequence discriminative training using data corresponding to that locale. The entire network is allowed to train in this stage. We find this further improves WER, see Table 1, resulting in grapheme-based models that significantly outperform the phoneme-based models.

Model	EnUs	EnGb	EnIn	EnAu
CTC-P	10.5	11.6	22.3	13.2
HCTC-G	11.2 (+6.7)	12.8 (+10.3)	22.9 (+2.6)	13.3 (+0.7)
Multi-dialect HCTC-G	10.1 (-3.8)	11.5 (-0.8)	22.6 (+1.3)	11.5 (-12.8)
Adapted multi-dialect HCTC-G	9.5 (-9.5)	10.5 (-9.5)	19.7 (-11.6)	11.0 (-16.7)

Table 1: WER (%) comparison of various grapheme-based models (rows 3-5) to phoneme-based models (row 2) for different locales. For each grapheme-based model a relative WER change compared to the corresponding phoneme model is also shown.

For a more fair comparison of phoneme and grapheme models given 12 million utterances, we trained phoneme models in a similar method on the multi-dialect data. Another method is to combine the phoneme sets for the four locales and train a single model on the combined data set, but this experiment is left for future investigation. In this paper, we train a single network with four phoneme CTC output layers with the 12 million multi-dialect data. This is similar to the network described in Section 2.5 but without the extra grapheme layers and grapheme CTC loss. Finally, the model is adapted to a specific locale with sequence discriminative training

and the non-matching phoneme losses are discarded. Table 2 shows the improvements in the phoneme system due to training on multi-dialect data.

Model	EnUs	EnGb	EnIn	EnAu
CTC-P	10.5	11.6	22.3	13.2
Adapted multi-dialect CTC-P	10.2	10.4	21.0	11.0

Table 2: WER (%) difference between training a phoneme-based model on 3 million utterances (top-row) and 12 million multi-accent utterances (bottom-row) for various locales.

We find that given single dialect 3 million utterances the phoneme-based models significantly outperform the grapheme-based models. However, with 12 million multi-dialect data the grapheme models begin to outperform or match the phoneme models, see Table 3. We believe that the grapheme-based models can learn a better model for pronunciations of the words from the acoustic data than conventional pronunciation models, but they can suffer from the data sparsity. Hence, they improve the recognition accuracy over the phoneme-based models with increasing amounts of data. Learning the phonetic transcriptions during training the grapheme-based models in multi-task learning is an effective way of alleviating the data sparsity problem.

Graphemes vs Phonemes				
Data	EnUs	EnGb	EnIn	EnAu
3 million single dialect	+6.7	+10.3	+2.6	+0.7
12 million multi dialect	-6.9	+0.9	-6.2	0.0

Table 3: Relative difference in WER (%) from a grapheme-based model compared to a phoneme-based model when both are trained with 3 million utterances (top-row) or 12 million utterance (bottom-row) for various locales.

4.3. Accented Speech

The multi-dialect training approach improves speech recognition performance significantly for accented speech. For Indian accented US traffic we see a reduction in WER from 15.2% to 11.2%, Table 4. Even more impressive gains are seen in the grapheme-based systems at a WER of 8.5% and even after the adaptation step with EnUs data (American accent) most of the accent robustness is retained with a WER of 8.7% (a relative improvement of 42%). The specific improvements for accented speech in the grapheme system compared to the phoneme system show the benefits of learning a pronunciation model directly from acoustic data. Since the phoneme-based systems rely on human transcribed pronunciations for a single accent, they are not easily adapted to different accents. We expect similar improvements for other combinations of accented speech, for example, British accent in India or Australian accent in the US and so on, however, we lack test sets to run evaluations.

Model	Indian Accent (US)
EnUs CTC-P	15.2
EnUs Adapted Multi-Dialect CTC-P	11.2
Multi-Dialect HCTC-G	8.5
EnUs Adapted Multi-Dialect HCTC-G	8.7

Table 4: WER (%) performance of various models on an Indian accented US queries test set.

4.4. Unseen Dialects

We evaluate the multi-dialect grapheme model on some unseen English dialects: South African (EnZa), Kenyan (EnKe), Nigerian (EnNg), Filipino (EnPh). We compare the performance to matched phoneme models trained specifically for those locales in Table 5. The multi-dialect grapheme model performs remarkably well for these unseen dialects and may perform as a generic English recognizer.

Model	EnZa	EnKe	EnNg	EnPh
Matched CTC-P	17.9	19.9	29.1	18.2
Multi-Dialect HCTC-G	16.7	21.0	32.6	19.7

Table 5: WER (%) performance of a phonene model trained specifically for a locale (top-row) and multi-dialect grapheme model that was not trained for that locale. WER is shown for South African (EnZa), Kenyan (EnKe), Nigerian (EnNg) and Filipino (EnPh) English.

5. FUTURE WORK: LIMITATIONS OF A GRAPHEME-BASED ACOUSTIC MODEL

One limitation of grapheme-based acoustic model is the data sparsity in the acoustic training data. Tail words with irregular pronunciations will not be learned by the model since there will not be many instances in the training data. An example is the word *Bexar* which is a county in Texas. Since, it is pronounced like *bear* and there are very small number of examples in training data, the grapheme-based system outputs the grapheme sequence, $b\ e\ a\ r$ and not $b\ e\ x\ a\ r$. This problem can easily be fixed in a phoneme-based system by adding the pronunciation of *Bexar* to the pronunciation dictionary. To deal with these irregularly pronounced tail words with the grapheme-based systems, we propose including a similar dictionary for graphemes that contains graphemic spellings for such words. For *Bexar*, such a dictionary would include a mapping to $b\ e\ a\ r$ and then the language model would choose *Bexar* appropriately just like in a phoneme-based system. We leave the impact of such a dictionary to future work.

6. CONCLUSIONS

We train grapheme-based acoustic models using a hierarchical LSTM-RNN architecture with phoneme and grapheme CTC multi-task losses. The grapheme-based system performs relatively 0.7% - 10.3% worse than a phoneme-based system on the tested dialects of English: US, British, Indian and Australian. However, when we train a single grapheme-based model with the combined data for all four dialects, we see significant improvements. The grapheme-based acoustic model implicitly learns a pronunciation model and thus benefits from additional acoustic training data. The combined multi-dialect grapheme-based model outperforms the dialect-specific phoneme-based models for all dialects except for Indian English. Finally, we adapt the multi-dialect grapheme-based model and see further improvements where they are 9.5% - 16.7% relative better than their phoneme-based counterparts. We train the phoneme recognizers in a similar fashion with the multi-dialect data and see relatively smaller improvements. With 3 million single-dialect data phoneme-based models perform better but with 12 million multi-dialect data grapheme-based models are better indicating that grapheme-based recognition might become more viable as larger data sets become available.

7. REFERENCES

[1] H. Bourlard and N. Morgan, *Connectionist speech recognition*, Kluwer Academic Publishers, 1994.

[2] C. Chelba and F. Jelinek, "Structured language modeling for speech recognition," in *Proceedings of NLDB99*, 1999.

[3] G. Alex, J. Navdeep, and M. Abdel-rahman, "Hybrid speech recognition with deep bidirectional LSTM," in *Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on*. IEEE, 2013, pp. 273–278.

[4] H. Sak, A. Senior, and F. Beaufays, "Long Short-Term Memory Recurrent Neural Network Architectures for Large Scale Acoustic Modeling," in *Interspeech*, 2014.

[5] H. Sak, O. Vinyals, G. Heigold, A. Senior, E. McDermott, R. Monga, and M. Mao, "Sequence discriminative distributed training of long short-term memory recurrent neural networks," in *Interspeech*, 2014.

[6] H. Sak, A. Senior, and F. Beaufays, "Long Short-Term Memory Based Recurrent Neural Network Architectures for Large Vocabulary Speech Recognition," *ArXiv e-prints*, Feb. 2014.

[7] H. Hermansky, D.P.W. Ellis, and S. Sharma, "Tandem connectionist feature extraction for conventional hmm systems," in *ICASSP*, 2000.

[8] G. Hinton, L. Deng, D. Yu, G. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition," *IEEE Signal Processing Magazine*, vol. 29, pp. 82–97, 2012.

[9] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9(8), pp. 17351780, 1997.

[10] T. Mikolov, M. Karafiat, J. Cernocky, , and S. Khudanpur, "Recurrent neural network based language model," in *Interspeech*, 2015.

[11] T. Mikolov, S. Kombrink, L. Burget, J. Cernocky, , and S. Khudanpur, "Extensions of recurrent neural network based language model," in *ICASSP*, 2011.

[12] T. Mikolov, A. Deoras, S. Kombrink, L. Burget, , and J. Cernocky, "Empirical evaluation and combination of advanced language modeling techniques," in *Interspeech*, 2011.

[13] K. Rao, F. Peng, H. Sak, and F. Beaufays, "Grapheme-to-phoneme conversion using long short-term memory recurrent neural networks," in *Proceedings of ICASSP*, 2015.

[14] I. Mcgraw, I. Badr, J. R. Glass, and Senior Member, "Learning lexicons from speech using a pronunciation mixture model," in *IEEE Transactions on Audio, Speech, and Language Processing*, 2013, vol. 21, p. 357366.

[15] L. Lu, A. Ghoshal, and S. Renals, "Acoustic data-driven pronunciation lexicon for large vocabulary speech recognition," in *Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on*, Dec 2013, pp. 374–379.

[16] A. Graves and N. Jaitly, "Towards end-to-end speech recognition with recurrent neural networks," in *Proceedings of ICML*, 2014.

[17] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend, and spell," in *arXiv preprint arXiv:1508.01211*, 2015.

[18] D. Amodei, R. Anubhai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, J. Chen, M. Chrzanowski, A. Coates, G. Diamos, E. Elsen, J. Engel, L. Fan, C. Fougner, T. Han, A. Hannun, B. Jun, P. LeGresley, L. Lin, S. Narang, A. Ng, S. Ozair, R. Prenger, J. Raiman, S. Satheesh, D. Seetapun, S. Sengupta, Y. Wang, Z. Wang, C. Wang, B. Xiao, D. Yogatama, J. Zhan, and Z. Zhu, "Deep speech 2: End-to-end speech recognition in english and mandarin," in *arXiv preprint arXiv:1512.02595*, 2015.

[19] F. Eyben, M. Wollmer, B. Schuller, and A. Graves, "From speech to letters using a novel neural network architecture for grapheme based asr," in *Automatic Speech Recognition and Understanding, ASRU*, 2009.

[20] Y. Sung, T. Hughes, F. Beaufays, and B. Strope, "Revisiting graphemes with increasing amounts of data," in *ICASSP*, 2008.

[21] S. Kanthak and H. Ney, "Multilingual acoustic modeling using graphemes," in *Proceedings of European Conference On Speech Communication and Tenchology*, 2003, pp. 1145–1148.

[22] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of ICML*, 2006.

[23] S. Fernández, A. Graves, and J. Schmidhuber, "Sequence labelling in structured domains with hierarchical recurrent neural networks.," in *Proceedings of the 20th International Joint Conference on Artificial Intelligence, IJCAI*, 2007.

[24] H. Sak, A. Senior, K. Rao, and F. Beaufays, "Fast and accurate recurrent neural network acoustic models for speech recognition," in *Interspeech*, 2015.

[25] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, Feb. 1989.

[26] H. Sak, F. Beaufays, K. Nakajima, and C. Allauzen, "Language model verbalization for automaticspeech recognition," in *ICASSP*, 2013.

[27] O. Siohan, "Sequence training of multi-task acoustic models using meta-state labels," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2016, pp. 5425–5429.

[28] M. L. Seltzer and J. Droppo, "Multi-task learning in deep neural networks for improved phoneme recognition," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, May 2013, pp. 6965–6969.

[29] G. Heigold, V. Vanhoucke, A. Senior, P. Nguyen, M. Ranzato, M. Devin, and J. Dean, "Multilingual acoustic models using distributed deep neural networks," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, May 2013, pp. 8619–8623.

[30] T. Fraga-Silva, J. L. Gauvain, and L. Lamel, "Speech recognition of multiple accented english data using acoustic model interpolation," in *2014 22nd European Signal Processing Conference (EUSIPCO)*, Sept 2014, pp. 1781–1785.

[31] B. Kingsbury, "Lattice-based optimization of sequence classification criteria for neural-network acoustic modeling," in *ICASSP*, Taipei, Taiwan, Apr. 2009, pp. 3761–3764.