PHONE SEQUENCE MODELING WITH RECURRENT NEURAL NETWORKS

*Nicolas Boulanger-Lewandowski**

Université de Montréal
Department IRO
Montréal, QC, Canada

Jasha Droppo *Mike Seltzer* *Dong Yu*

Microsoft Research
One Microsoft Way
Redmond, WA, USA

ABSTRACT

In this paper, we investigate phone sequence modeling with recurrent neural networks in the context of speech recognition. We introduce a hybrid architecture that combines a phonetic model with an arbitrary frame-level acoustic model and we propose efficient algorithms for training, decoding and sequence alignment. We evaluate the advantage of our phonetic model on the TIMIT and Switchboard-mini datasets in complementarity to a powerful context-dependent deep neural network (DNN) acoustic classifier and a higher-level 3-gram language model. Consistent improvements of 2–10% in phone accuracy and 3% in word error rate suggest that our approach can readily replace HMMs in current state-of-the-art systems.

***Index Terms*—** Recurrent neural network, phonetic model, speech recognition

1. INTRODUCTION

Automatic speech recognition is an active area of research in the signal processing and machine learning communities [1]. Existing approaches are commonly based on three fundamental modules: (1) an *acoustic* model that focuses on the discriminative aspect of the audio signal, (2) a *phonetic* model that attempts to describe the temporal dependencies associated with the sequence of phone labels, and (3) a *language* model that describes the higher-level dependencies between words and sentences. In this work, we wish to replace the popular hidden Markov model (HMM) approach with a more powerful neural network-based phonetic model.

Recurrent neural networks (RNN) [2] are powerful dynamical systems that incorporate an internal memory, or *hidden state*, represented by a self-connected layer of neurons. This property makes them well suited to model temporal sequences, such as frames in a magnitude spectrogram or phone labels in a spoken utterance, by being trained to predict the output at the next time step given the previous ones. RNNs are completely general in that in principle they can describe arbitrarily complex long-term temporal dependencies, which made them very successful in music and language applications [3, 4, 5].

While RNN-based language models significantly surpass popular alternatives like HMMs, it is not immediately obvious how to combine the acoustic and phonetic models under a single training objective. The simple approach of multiplying the predictions of both models before renormalizing as in a maximum-entropy Markov chain [6] often results in the so-called *label bias problem* where the symbolic information overwhelms the acoustic information in low-entropy sequences with frequently reoccuring symbols [7]. Several attempts have been made to reduce those difficulties, such as with conditional random fields [7], regularization of the symbolic and acoustic sources [8], by increasing the entropy per time step with a lower temporal resolution [9], modeling *unaligned* phonetic sequences with an implicit exponential duration model [10, 11], or with the popular approach of stacking an HMM on top of a frame-level classifier (e.g. [12]). In this paper, we propose an alternative approach that enforces a proper weighting of the acoustic and symbolic predictors and allows the probability flow of a candidate solution to vary according to the acoustic observations [7]. Our hybrid architecture is a generative model that generalizes the HMM and that can be trained similarly following the expectation-maximization principle while exploiting the predictive power of RNNs in describing complex temporal dependencies. An advantage of our design not present in [11] is the possibility of leveraging *arbitrary* frame-level acoustic classifiers such as a DNN trained with dropout or advanced optimization techniques [13]. We also propose efficient inference algorithms for decoding and optimal sequence alignment inspired from Viterbi decoding. Finally, we investigate the extent to which phone sequence modeling is relevant in complementarity to powerful context-dependent acoustic classifiers and higher-level language models.

The remainder of the paper is organized as follows. In sections 2 and 3 we introduce the RNN architecture and our hybrid phone sequence model. In sections 4 and 5 we detail our decoding and alignment algorithms. Finally, we present our methodology and results in section 6.

2. RECURRENT NEURAL NETWORKS

The RNN formally defines the distribution of the output sequence $z \equiv \{z^{(t)} \in C, t \leq T\}$ of length T, where C is the dictionary of possible phone labels ($|C| = N$):

$$P(z) = \prod_{t=1}^{T} P(z^{(t)}|\mathcal{A}^{(t)}) \tag{1}$$

where $\mathcal{A}^{(t)} \equiv \{z^{(\tau)}|\tau < t\}$ is the sequence history at time t, and $P(z^{(t)}|\mathcal{A}^{(t)})$ is the conditional probability of observing $z^{(t)}$ according to the model, defined below in equation (5).

A single-layer RNN with hidden units $h^{(t)}$ is defined by its recurrence relation:

$$h^{(t)} = \sigma(W_{zh}z^{(t-1)} + W_{hh}h^{(t-1)} + b_h) \tag{2}$$

where the indices of weight matrices and bias vectors have obvious meanings.

*This investigation was carried out during his internship at Microsoft Research, Redmond.

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

The prediction $y^{(t)}$ is obtained from the hidden units at the current time step $h^{(t)}$ and the previous output $z^{(t-1)}$:

$$y^{(t)} = s(W_{hz}h^{(t)} + W_{zz}z^{(t-1)} + b_z) \tag{3}$$

where the W_{zz} matrix is useful to explicitly disallow certain state transitions by setting the corresponding entries to very large negative values, and $s(a)$ is the softmax function of an activation vector a:

$$(s(a))_j \equiv \frac{\exp(a_j)}{\sum_{j'=1}^{N}\exp(a_{j'})}, \tag{4}$$

and should be as close as possible to the target vector $z^{(t)}$. In the case of multiclass classification problems such as frame-level phone recognition, the target is a one-hot vector and the likelihood of an observation is given by the dot product:

$$P(z^{(t)}|\mathcal{A}^{(t)}) = z^{(t)} \cdot y^{(t)}. \tag{5}$$

The RNN model can be trained by maximum likelihood with the cross-entropy cost:

$$L(z) = -\sum_{t=1}^{T}\log(z^{(t)} \cdot y^{(t)}) \tag{6}$$

where the gradient with respect to the model parameters is obtained by backpropagation through time (BPTT) [2].

While in principle a properly trained RNN can describe arbitrarily complex temporal dependencies at multiple time scales, in practice gradient-based training suffers from various pathologies [14]. Several strategies can be used to help reduce these difficulties including gradient clipping, leaky integration, sparsity and Nesterov momentum [5].

3. PHONE SEQUENCE MODELING

In this section, we generalize the popular technique of superposing an HMM to an acoustic model by replacing the HMM with an arbitrary phonetic model. This will allow to exploit the power of RNNs for phone modeling while providing a principled way to combine the two models.

Our hybrid acoustic-phonetic sequence model is a graphical model composed of an underlying phone sequence z:

$$P(z^{(t)}|\{x^{(\tau)}, \tau < t\}, \mathcal{A}^{(t)}) = P(z^{(t)}|\mathcal{A}^{(t)}) \tag{7}$$

and an acoustic sequence x *emitted* given the phone sequence:

$$P(x^{(t)}|\{x^{(\tau)}, \tau \neq t\}, z) = P(x^{(t)}|z^{(t)}). \tag{8}$$

The emission probability (8) can be reformulated using Bayes' rule [15]:

$$P(x^{(t)}|z^{(t)}) \propto \frac{P(z^{(t)}|x^{(t)})}{P(z^{(t)})} \tag{9}$$

where $P(z^{(t)}|x^{(t)})$ is the output of an acoustic classifier, $P(z^{(t)})$ is the marginal distribution of phones and constant terms given x have been removed. This adjustment is referred to as scaled likelihood estimation in [12].

The next-step phone sequence distribution has a general expression in the right-hand side of equation (7) to accomodate different phonetic models. For an HMM, this distribution depends only on $z^{(t-1)}$:

$$P(z^{(t)} = i|\mathcal{A}^{(t)}) = \begin{cases} T_{z^{(t-1)},i} & \text{if } t > 0 \\ \pi_i & \text{if } t = 0 \end{cases} \tag{10}$$

where $T_{j,i}$ is the row-normalized transition matrix and π_i the initial occupancy of phone i. In our case, we will replace (10) with the distribution of an RNN (eq. 5) which depends on the full sequence history $\mathcal{A}^{(t)}$.

By combining equations (7)-(9), we obtain the conditional distribution over phones z given the input x:

$$P(z|x) = \prod_{t=1}^{T} P(z^{(t)}|x^{(t)}, \mathcal{A}^{(t)}) \tag{11}$$

$$P(z^{(t)}|x^{(t)}, \mathcal{A}^{(t)}) \propto \frac{P(z^{(t)}|x^{(t)})}{P(z^{(t)})} P(z^{(t)}|\mathcal{A}^{(t)}) \tag{12}$$

which can be interpreted as the output of the hybrid model.

As argued previously [16], a limitation of the hybrid model occurs when the acoustic model has access to contextual information when predicting $z^{(t)}$, either directly in the form of an input window around $x^{(t)}$ or indirectly via the hidden state of an RNN. When the input includes information from neighboring frames, the independence assumption (8) breaks down, making it difficult to combine the two models in equation (12). Intuitively, multiplying the predictions $P(z^{(t)}|x^{(t)})$ and $P(z^{(t)}|\mathcal{A}^{(t)})$ to estimate the joint distribution will count certain factors twice since both models have been trained separately. Note that the marginals $P(z^{(t)})$ are counted only once with scaled likelihood estimation (eq. 9), but it is reasonable to expect that certain temporal dependencies will be captured by both models. In our experiments, we found that this conceptual difficulty surprisingly did not prevent good performance. Furthermore, the alternative approach of multiplying the two predictions and renormalizing in order to train the system jointly [17, 10, 11] suffered heavily from the label bias problem, and we found it crucial not to renormalize the two distributions to achieve good performance. Note that the transducer approach in [11] circumvents the label bias problem by modeling *unaligned* phone sequences with an implicit exponential duration model.

During training, we wish to maximize the log-likelihood $\log P(x, z)$ of training example pairs x, z. It is easy to see from equations (11) and (12) that a stochastic gradient ascent update involves terms associated with the phonetic and acoustic models that can be computed separately:

$$\frac{\partial \log P(x,z)}{\partial \Theta_a} = \frac{\partial}{\partial \Theta_a}\sum_{t=1}^{T}\log P(z^{(t)}|x^{(t)}) \tag{13}$$

$$\frac{\partial \log P(x,z)}{\partial \Theta_p} = \frac{\partial}{\partial \Theta_p}\sum_{t=1}^{T}\log P(z^{(t)}|\mathcal{A}^{(t)}) \tag{14}$$

where Θ_a, Θ_p denote the parameters of the acoustic and phonetic models respectively.

When only unaligned phone sequences $\bar{z} \equiv \{\bar{z}^{(u)}, u \leq U\}$ of length U are available during training, the hard expectation-maximization (EM) approach can be adopted, by regarding the alignments as missing data. After initializing the aligned sequences z from a flat start or another existing method, we alternate updates to the model parameters (M step) and to the estimated alignments given the current parameters (E step) as described in section 5. Both of these steps are guaranteed to increase the training objective $\log P(x, z)$ unless a local maximum is already reached.

4. DECODING

In our architecture, the phonetic model implicitly ties $z^{(t)}$ to its history $\mathcal{A}^{(t)}$ and encourages coherence between successive output

frames, and temporal smoothing in particular. At test time, predicting one time step $z^{(t)}$ requires the knowledge of the previous decisions on $z^{(\tau)}$ (for $\tau < t$) which are yet uncertain (not chosen optimally), and proceeding in a greedy chronological manner does not necessarily yield configurations that maximize the likelihood of the complete sequence. We rather favor a global search approach analogous to the Viterbi algorithm for discrete-state HMMs to infer the sequence $z^* \equiv \{z^{(t)*} | t \leq T\}$ with maximal probability given the input.

For HMM phonetic models, the distribution in equation (12) becomes:

$$P(z^{(t)}|x^{(t)}, \mathcal{A}^{(t)}) \propto \frac{P(z^{(t)}|x^{(t)})}{P(z^{(t)})} P(z^{(t)}|z^{(t-1)}). \quad (15)$$

Since it depends only on $z^{(t-1)}$, it is easy to derive a recurrence relation to optimize z^* by dynamic programming, giving rise to the well-known Viterbi algorithm.

The inference algorithm we propose for RNN phonetic models is based on a dynamic programming-like (DP) pruned beam search introduced in [9]. Beam search is a breadth-first tree search where only the w most promising paths (or nodes) at depth t are kept for future examination. In our case, a node at depth t corresponds to a subsequence of length t, and all descendants of that node are assumed to share the same sequence history $\mathcal{A}^{(t+1)}$. Note that $w = 1$ reduces to a greedy search, and $w = N^T$ corresponds to an exhaustive breadth-first search.

A pathological condition that sometimes occurs with beam search is the exponential duplication of highly likely quasi-identical paths differing only at a few time steps, that quickly saturate beam width with essentially useless variations. A natural extension to beam search is to make a better use of the available width w via pruning. A particularly efficient pruning strategy is to consider only the most promising path out of all partial paths with identical $z^{(t)}$ when making a decision at time t. This leads to the solution of keeping track of the N most likely paths arriving at each possible label $j \in C$ with the recurrence relations:

$$l_j^{(t)} = l_{k_j^{(t)}}^{(t-1)} + P(z^{(t)} = j | x, s_{k_j^{(t)}}^{(t-1)}) \quad (16)$$

$$s_j^{(t)} = \{s_{k_j^{(t)}}^{(t-1)}, j\} \quad (17)$$

$$\text{with } k_j^{(t)} \equiv \underset{k=1}{\overset{N}{\operatorname{argmax}}} \left[l_k^{(t-1)} + P(z^{(t)} = j | x, s_k^{(t-1)}) \right] \quad (18)$$

and initial conditions $l_j^{(0)} = 0, s_j^{(0)} = \{\}$, where the variables $l_j^{(t)}, s_j^{(t)}$ represent respectively the maximal cumulative log-likelihood and the associated partial output sequence ending with label j at time t [9]. In our case, $P(z^{(t)} = j | x, s_k^{(t-1)})$ is given by equation (12): since the acoustic prediction and the marginal distribution do not depend on $\mathcal{A}^{(t)}$, we can compute those contributions in advance.

It should not be misconstrued that the algorithm is limited to "local" or greedy decisions for two reasons: (1) the complete sequence history $\mathcal{A}^{(t)}$ is relevant for the prediction $y^{(t)}$ at time t, and (2) a decision $z^{(t)*}$ at time t can be affected by an observation $x^{(t+\delta t)}$ arbitrarily far in the future via *backtracking*, analogously to Viterbi decoding.

5. OPTIMAL ALIGNMENT

In this section, we propose an algorithm to search for the aligned phone sequence $z \equiv \{z^{(t)} | t \leq T\}$ with maximal probability $P(z|x)$ according to a trained model (eq. 11), that is consistent with a given unaligned phone sequence $\bar{z} \equiv \{\bar{z}^{(u)} | u \leq U\}$ where $U < T$. The sequences z and $\bar{z}$ are said to be consistent if there exists an alignment $a \equiv \{u_t | t \leq T\}$ satisfying $u_1 = 1$, $u_T = U$ and $u_t - u_{t-1} \in \{0, 1\}$ for which $z^{(t)} = \bar{z}^{(u_t)}, \forall t \leq T$. The objective is to find the optimal alignment a^*.

Since an exact solution is intractable in the general case that the predictions fully depend on the sequence history, we hypothesize that it is sufficient to consider only the most promising path out of all partial paths with identical u_t when making a decision at time t.[1] Under this assumption, any subsequence $\{u_t^* | t \leq T'\}$ of the global optimum $\{u_t^* | t \leq T\}$ ending at time $T' < T$ must also be optimal under the constraint $u_{T'} = u_{T'}^*$. This last constraint is necessary to avoid a greedy solution. Setting $T' = T - 1$ leads to the DP-like solution of keeping track of the (at most) U most likely paths arriving at each possible index u, $\max(1, U - T + t) \leq u \leq \min(U, t)$ with the recurrence relations:

$$l_u^{(t)} = l_{k_u^{(t)}}^{(t-1)} + P(z^{(t)} = \bar{z}^{(u)} | x, s_{k_u^{(t)}}^{(t-1)}) \quad (19)$$

$$s_u^{(t)} = \{s_{k_u^{(t)}}^{(t-1)}, \bar{z}^{(u)}\} \quad (20)$$

$$\text{with } k_u^{(t)} \equiv \underset{k \in \{u-1, u\}}{\operatorname{argmax}} \left[l_k^{(t-1)} + P(z^{(t)} = \bar{z}^{(u)} | x, s_k^{(t-1)}) \right] \quad (21)$$

and initial conditions $l_u^{(0)} = 0, s_u^{(0)} = \{\}$, where the variables $l_j^{(t)}, s_j^{(t)}$ are defined similarly as in equations (16)-(18). The optimal aligned sequence is then given by $z^* \simeq s_U^{(T)}$. This algorithm has a time complexity $O(TU)$ independent of N.

Since finding an optimal alignment in the inner loop of an EM iteration can be prohibitive, we can further postulate that the optimal alignment a^* is close to an approximate alignment a' that can be computed much more cheaply. Typically, a' would be obtained by an acoustic model whose predictions depend only on x, eliminating the need to maintain the hidden states of multiple RNNs. Assuming that the distance between a^* and a' is δ:

$$|a^*, a'| \equiv \max_{t=1}^{T} |u_t^* - u_t'| = \delta, \quad (22)$$

the range of plausible values for u can be significantly reduced in equations (19)-(21). Values of δ as low as 2–4 were found to work well in practice, producing identical alignments in a majority of cases with less than 10% of the computation.

6. EXPERIMENTS

In this section, we evaluate the performance of our RNN phonetic model and hybrid training procedure relatively to a baseline HMM system. We use two datasets to evaluate our method: the TIMIT corpus and the 30 hour "mini-train" subset of the Switchboard corpus. We report phone accuracy on the TIMIT data, which includes expertly-annotated phone sequences. We report phone accuracy and word accuracy on the Switchboard data, where the correct phonetic transcription is approximated by a dictionary-based alignment of the data by our baseline DNN + HMM system.

The TIMIT experiments rely on a 123 dimensional acoustic feature vector, calculated as 40 dimensional mel-frequency log-filterbank features, together with an energy measure and first and

[1] Replacing the pruning condition on u_t with a condition on $z^{(t)}$ as for decoding is not as effective because $\tilde{u}_t \neq \hat{u}_t, \tilde{z}^{(t)} = \hat{z}^{(t)}$ for two candidates $\tilde{a}, \hat{a}$ indicate fundamentally different alignments.

Acoustic model	HMM (dev)	HMM (test)	RNN (dev)	RNN (test)	Hybrid (dev)	Hybrid (test)
LR	62.6	61.8	63.8	62.8	65.3	63.5
RNN	69.9	68.6	70.6	69.4	74.2	72.2
DNN	79.0	77.1	79.8	77.9	**80.4**	**78.6**

Table 1. Development and test phone accuracies (%) obtained on the TIMIT dataset using different combinations of acoustic and phonetic models.

Acoustic model	HMM	RNN	Hybrid
LR	31.8	32.3	34.4
RNN	40.5	43.8	44.7
DNN	70.0	72.7	**73.7**

Table 2. Development phone accuracies (%) obtained on the Switchboard dataset using different combinations of acoustic and phonetic models.

second temporal derivatives. The Switchboard experiments use a 52 dimensional acoustic feature vector, consisting of a basic 13-dimensional PLP cepstral vector together with its first, second, and third temporal derivatives.

We consider three acoustic models: a simple logistic regression (LR) classifier, an RNN using x as input (replacing $z^{(t-1)}$ in eq. 2) and a DNN with 4×1024 (TIMIT) or 5×2048 (Switchboard) hidden units trained with context-dependent triphones. The DNN features are the activations of the final hidden layer of the fully trained model. For each acoustic model, we compare three phonetic models: an HMM baseline, an RNN trained with fixed baseline alignments, and an RNN trained with our hybrid EM procedure. Early stopping is performed based on the cross-entropy of a held-out development set, which was randomly selected from 5% of the training set for Switchboard. The phone accuracy is determined as:

$$PA = 1 - \frac{\sum_{\bar{z},\bar{z}_0} L(\bar{z}, \bar{z}_0)}{\sum_{\bar{z}_0} |\bar{z}_0|} \tag{23}$$

where $L(\cdot, \cdot)$ is the Levenshtein distance between two sequences and $\bar{z}, \bar{z}_0$ represent respectively the predicted and ground-truth sequences.

Developement and test phone accuracies are presented for the two datasets in Tables 1 and 2 for different combinations of acoustic and phonetic models. We observe consistent improvements with the RNN phonetic model, especially when trained using the hybrid procedure, attaining accuracies between 2–10% over the baseline. Note that the improvements obtained with CRF full-sequence training are typically more modest in this context [18], suggesting that Markovian assumptions in linear-chain CRFs are more limiting than the conditional independence assumption violated by our model as discussed in section 3.

It could be argued that the improvements brought by our RNN phonetic model capitalize on higher-level dependencies between phones, and that the inclusion of a word language model would nullify those gains. In the next experiments we verify if our method translates in good word recognition performance on the Switchboard dataset. While a 3-gram language model could be directly integrated into a sophisticated context-dependent decoding procedure, we simply provide a performance benchmark by *rescoring* a list of the N best candidates found by a DNN + HMM system ($N = 100$). The

DNN + HMM	33.0
DNN + RNN	32.7
DNN (Hybrid)	**32.0**
Oracle	19.5
Anti-oracle	56.8

Table 3. Test word error rates (%) obtained on the Switchboard dataset using different phonetic models.

word error rates shown in Table 3 clearly demonstrate the superiority of an RNN phonetic model when used in complementarity to a language model.

7. CONCLUSIONS

In this paper, we presented a principled way to combine an RNN-based phonetic model with an arbitrary frame-level acoustic classifier. The efficiency of the decoding and alignment procedures now allows to use an RNN whenever an HMM was previously used. Interestingly, phone sequence modeling seems to be an important component of accurate speech recognition, even in the case where strong acoustic classifiers and word language models are already available.

8. REFERENCES

[1] J. Baker, L. Deng, J. Glass, S. Khudanpur, C.-H. Lee, N. Morgan, and D. O'Shaughnessy, "Developments and directions in speech recognition and understanding," *IEEE Signal Processing Magazine*, vol. 26, no. 3, pp. 75–80, 2009.

[2] D. E. Rumelhart, G. Hinton, and R. J. Williams, "Learning internal representations by error propagation," in *Parallel Dist. Proc.*, pp. 318–362. MIT Press, 1986.

[3] N. Boulanger-Lewandowski, Y. Bengio, and P. Vincent, "Modeling temporal dependencies in high-dimensional sequences: Application to polyphonic music generation and transcription," in *ICML 29*, 2012.

[4] T. Mikolov, A. Deoras, S. Kombrink, L. Burget, and J. Cernocký, "Empirical evaluation and combination of advanced language modeling techniques.," in *INTERSPEECH*, 2011, pp. 605–608.

[5] Y. Bengio, N. Boulanger-Lewandowski, and R. Pascanu, "Advances in optimizing recurrent networks," in *ICASSP*, 2013.

[6] A. McCallum, D. Freitag, and F. Pereira, "Maximum entropy markov models for information extraction and segmentation.," in *ICML 17*, 2000, pp. 591–598.

[7] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *ICML 18*, 2001, pp. 282–289.

[8] N. Boulanger-Lewandowski, Y. Bengio, and P. Vincent, "High-dimensional sequence transduction," in *ICASSP*, 2013.

[9] N. Boulanger-Lewandowski, Y. Bengio, and P. Vincent, "Audio chord recognition with recurrent neural networks," in *ISMIR*, 2013.

[10] A. Graves, "Sequence transduction with recurrent neural networks," in *ICML 29*, 2012.

[11] A. Graves, A.-R Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *ICASSP*, 2013.

[12] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 30–42, 2012.

[13] G. E. Dahl, T. Sainath, and G. Hinton, "Improving deep neural networks for LVCSR using rectified linear units and dropout," in *ICASSP*, 2013.

[14] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE Trans. on Neural Networks*, vol. 5, no. 2, pp. 157–166, 1994.

[15] G. Hinton, L. Deng, D. Yu, G. E. 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," *Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[16] P. Brown, *The acoustic-modeling problem in automatic speech recognition*, Ph.D. thesis, Carnegie-Mellon University, 1987.

[17] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *ICML 23*, 2006, pp. 369–376.

[18] A.-R. Mohamed, D. Yu, and L. Deng, "Investigation of full-sequence training of deep belief networks for speech recognition.," in *INTERSPEECH*, 2010, pp. 2846–2849.