WAV2LETTER: AN END-TO-END CONVNET-BASED SPEECH RECOGNITION SYSTEM

Ronan Collobert
Facebook AI Research, Menlo Park
locronan@fb.com

Christian Puhrsch
Facebook AI Research, Menlo Park
cpuhrsch@fb.com

Gabriel Synnaeve
Facebook AI Research, New York
gab@fb.com

ABSTRACT

This paper presents a simple end-to-end model for speech recognition, combining a convolutional network based acoustic model and a graph decoding. It is trained to output letters, with transcribed speech, without the need for force alignment of phonemes. We introduce an automatic segmentation criterion for training from sequence annotation without alignment that is on par with CTC (Graves et al., 2006) while being simpler. We show competitive results in word error rate on the Librispeech corpus (Panayotov et al., 2015) with MFCC features, and promising results from raw waveform.

1 INTRODUCTION

We present an end-to-end system to speech recognition, going from the speech signal (e.g. Mel-Frequency Cepstral Coefficients (MFCC), power spectrum, or raw waveform) to the transcription. The acoustic model is trained using letters (graphemes) directly, which take out the need for an intermediate (human or automatic) phonetic transcription. Indeed, the classical pipeline to build state of the art systems for speech recognition consists in first training an HMM/GMM model to force align the units on which the final acoustic model operates (most often context-dependent phone states). This approach takes its roots in HMM/GMM training (Woodland & Young, 1993). The improvements brought by deep neural networks (DNNs) (Mohamed et al., 2012; Hinton et al., 2012) and convolutional neural networks (CNNs) (Sercu et al., 2015; Soltau et al., 2014) for acoustic modeling only extend this training pipeline.

The current state of the art on Librispeech (the dataset that we used for our evaluations) uses this approach too (Panayotov et al., 2015; Peddinti et al., 2015b), with an additional step of speaker adaptation (Saon et al., 2013; Peddinti et al., 2015a). Recently, Senior et al. (2014) proposed GMM-free training, but the approach still requires to generate a force alignment. An approach that cut ties with the HMM/GMM pipeline (and with force alignment) was to train with a recurrent neural network (RNN) (Graves et al., 2013) for phoneme transcription. There are now competitive end-to-end approaches of acoustic models toppled with RNNs layers as in (Hannun et al., 2014; Miao et al., 2015; Saon et al., 2015; Amodei et al., 2015), trained with a sequence criterion (Graves et al., 2006). However these models are computationally expensive, and thus take a long time to train.

Compared to classical approaches that need phonetic annotation (often derived from a phonetic dictionary, rules, and generative training), we propose to train the model end-to-end, using graphemes directly. Compared to sequence criterion based approaches that train directly from speech signal to graphemes (Miao et al., 2015), we propose a simple(r) architecture (23 millions of parameters for our best model, vs. 100 millions of parameters in (Amodei et al., 2015)) based on convolutional networks

for the acoustic model, toppled with a graph transformer network (Bottou et al., 1997), trained with a simpler sequence criterion. Our word-error-rate on clean speech is slightly better than (Hannun et al., 2014), and slightly worse than (Amodei et al., 2015), in particular factoring that they train on 12,000 hours while we only train on the 960h available in LibriSpeech's train set. Finally, some of our models are also trained on the raw waveform, as in (Palaz et al., 2013; 2015; Sainath et al., 2015). The rest of the paper is structured as follows: the next section presents the convolutional networks used for acoustic modeling, along with the automatic segmentation criterion. The following section shows experimental results comparing different features, the criterion, and our current best word error rates on LibriSpeech.

2 ARCHITECTURE

Our speech recognition system is a standard convolutional neural network (LeCun & Bengio, 1995) fed with various different features, trained through an alternative to the Connectionist Temporal Classification (CTC) (Graves et al., 2006), and coupled with a simple beam search decoder. In the following sub-sections, we detail each of these components.

2.1 FEATURES

We consider three types of input features for our model: MFCCs, power-spectrum, and raw wave. MFCCs are carefully designed speech-specific features, often found in classical HMM/GMM speech systems (Woodland & Young, 1993) because of their dimensionality compression (13 coefficients are often enough to span speech frequencies). Power-spectrum features are found in most recent deep learning acoustic modeling features (Amodei et al., 2015). Raw wave has been somewhat explored in few recent work (Palaz et al., 2013; 2015). ConvNets have the advantage to be flexible enough to be used with either of these input feature types. Our acoustic models output letter scores (one score per letter, given a dictionary $\mathcal{L}$).

2.2 CONVNET ACOUSTIC MODEL

The acoustic models we considered in this paper are all based on standard 1D convolutional neural networks (ConvNets). ConvNets interleave convolution operations with pointwise non-linearity operations. Often ConvNets also embark pooling layers: these type of layers allow the network to "see" a larger context, without increasing the number of parameters, by locally aggregating the previous convolution operation output. Instead, our networks leverage striding convolutions. Given $(x_t)_{t=1\dots T_x}$ an input sequence with T_x frames of d_x dimensional vectors, a convolution with kernel width kw, stride dw and d_y frame size output computes the following:

$$y_t^i = b_i + \sum_{j=1}^{d_x} \sum_{k=1}^{kw} w_{i,j,k}\, x^j_{dw\times(t-1)+k} \qquad \forall 1 \leq i \leq d_y, \qquad (1)$$

where $b \in \mathbb{R}^{d_y}$ and $w \in \mathbb{R}^{d_y \times d_x \times kw}$ are the parameters of the convolution (to be learned).

Pointwise non-linear layers are added after convolutional layers. In our experience, we surprisingly found that using hyperbolic tangents, their piecewise linear counterpart HardTanh (as in (Palaz et al., 2015)) or ReLU units lead to similar results.

There are some slight variations between the architectures, depending on the input features. MFCC-based networks need less striding, as standard MFCC filters are applied with large strides on the input

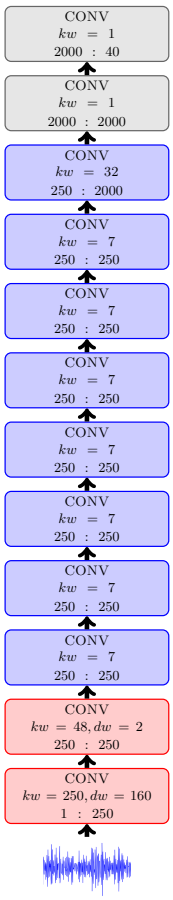


Figure 1: Our neural network architecture for raw wave. First two layers are convolutions with strides. Last two layers are convolutions with $kw = 1$, which are equivalent to fully connected layers. Power spectrum and MFCC based networks do not have the first layer.

raw sequence. With power spectrum-based and raw wave-based networks, we observed that the overall stride of the network was more important than where the convolution with strides were placed. We found thus preferrable to set the strided convolutions near the first input layers of the network, as it leads to the fastest architectures: with power spectrum features or raw wave, the input sequences are very long and the first convolutions are thus the most expensive ones.

The last layer of our convolutional network outputs one score per letter in the letter dictionary ($d_y = |\mathcal{L}|$). Our architecture for raw wave is shown in Figure 1 and is inspired by (Palaz et al., 2015). The architectures for both power spectrum and MFCC features do not include the first layer. The full network can be seen as a non-linear convolution, with a kernel width of size 31280 and stride equal to 320; given the sample rate of our data is 16KHz, label scores are produced using a window of 1955 ms, with steps of 20ms.

2.3 Inferring Segmentation with AutoSegCriterion

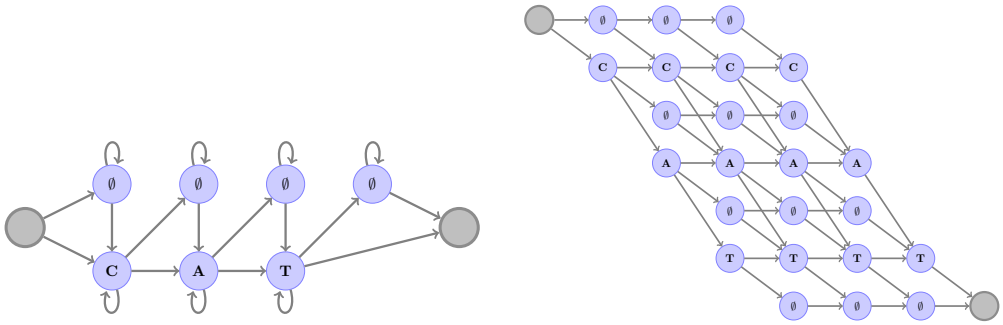
Most large labeled speech databases provide only a text transcription for each audio file. In a classification framework (and given our acoustic model produces letter predictions), one would need the segmentation of each letter in the transcription to train properly the model. Unfortunately, manually labeling the segmentation of each letter would be tedious. Several solutions have been explored in the speech community to alleviate this issue: HMM/GMM models use an iterative EM procedure: (i) during the Estimation step, the best segmentation is inferred, according to the current model, by maximizing the joint probability of the letter (or any sub-word unit) transcription and input sequence. (ii) During the Maximization step the model is optimized by minimizing a frame-level criterion, based on the (now fixed) inferred segmentation. This approach is also often used to boostrap the training of neural network-based acoustic models.

Other alternatives have been explored in the context of hybrid HMM/NN systems, such as the MMI criterion (Bahl et al., 1986) which maximizes the mutual information between the acoustic sequence and word sequences or the Minimum Bayse Risk (MBR) criterion (Gibson & Hain, 2006).

More recently, standalone neural network architectures have been trained using criterions which jointly infer the segmentation of the transcription while increase the overall score of the right transcription (Graves et al., 2006; Palaz et al., 2014). The most popular one is certainly the Connectionist Temporal Classification (CTC) criterion, which is at the core of Baidu's Deep Speech architecture (Amodei et al., 2015). CTC assumes that the network output probability scores, normalized at the frame level. It considers all possible sequence of letters (or any sub-word units), which can lead to a to a given transcription. CTC also allow a special "blank" state to be optionally inserted between each letters. The rational behind the blank state is two-folds: (i) modeling "garbage" frames which might occur between each letter and (ii) identifying the separation between two identical consecutive letters in a transcription. Figure 2a shows an example of the sequences accepted by CTC for a given transcription. In practice, this graph is unfolded as shown in Figure 2b, over the available frames output by the acoustic model. We denote $\mathcal{G}_{ctc}(\theta, T)$ an unfolded graph over T frames for a given transcription θ, and $\pi = \pi_1, \ldots, \pi_T \in \mathcal{G}_{ctc}(\theta, T)$ a path in this graph representing a (valid) sequence of letters for this transcription. At each time step t, each node of the graph is assigned with the corresponding log-probability letter (that we denote $f_t(\cdot)$) output by the acoustic model. CTC aims at maximizing the "overall" score of paths in $\mathcal{G}_{ctc}(\theta, T)$; for that purpose, it minimizes the Forward score:

$$CTC(\theta, T) = - \operatorname*{logadd}_{\pi \in \mathcal{G}_{ctc}(\theta, T)} \sum_{t=1}^{T} f_{\pi_t}(x)\,, \tag{2}$$

where the "logadd" operation, also often called "log-sum-exp" is defined as $\mathrm{logadd}(a, b) = \exp(\log(a) + \log(b))$. This overall score can be efficiently computed with the Forward algorithm. To put things in perspective, if one would replace the $\mathrm{logadd}(\cdot)$ by a $\max(\cdot)$ in (2) (which can be then efficiently computed by the Viterbi algorithm, the counterpart of the Forward algorithm), one would then maximize the score of the *best* path, according to the model belief. The $\mathrm{logadd}(\cdot)$ can be seen as a smooth version of the $\max(\cdot)$: paths with similar scores will be attributed the same weight in the overall score (and hence receive the same gradient), and paths with much larger score will have much more overall weight than paths with low scores. In practice, using the $\mathrm{logadd}(\cdot)$ works much better than the $\max(\cdot)$. It is also worth noting that maximizing (2) does not diverge, as the acoustic model is assumed to output normalized scores (log-probabilities) $f_i(\cdot)$.



(a) (b)

Figure 2: The CTC criterion graph. (a) Graph which represents all the acceptable sequences of letters (with the blank state denoted "$\emptyset$"), for the transcription "cat". (b) Shows the same graph unfolded over 5 frames. There are no transitions scores. At each time step, nodes are assigned a conditional probability output by the neural network acoustic model.

In this paper, we explore an alternative to CTC, with three differences: (i) there are no blank labels, (ii) un-normalized scores on the nodes (and possibly un-normalized transition scores on the edges) (iii) global normalization instead of per-frame normalization:

- The advantage of (i) is that it produces a much simpler graph (see Figure 3a and Figure 3b). We found that in practice there was no advantage of having a blank class to model the possible "garbage" frames between letters. Modeling letter repetitions (which is also an important quality of the blank label in CTC) can be easily replaced by repetition character labels (we used two extra labels for two and three repetitions). For example "caterpillar" could be written as "caterpil2ar", where "2" is a label to represent the repetition of the previous letter. Not having blank labels also simplifies the decoder.
- With (ii) one can easily plug an external language model, which would insert transition scores on the edges of the graph. This could be particularly useful in future work, if one wanted to model representations more high-level than letters. In that respect, avoiding normalized transitions is important to alleviate the problem of "label bias" Bottou (1991); Lafferty et al. (2001). In this work, we limited ourselves to transition scalars, which are learned together with the acoustic model.
- The normalization evoked in (iii) is necessary when using un-normalized scores on nodes or edges; it insures incorrect transcriptions will have a low confidence.

In the following, we name our criterion "Auto Segmentation Criterion" (ASG). Considering the same notations than for CTC in (2), and an unfolded graph $\mathcal{G}_{asg}(\theta, T)$ over T frames for a given transcription θ (as in Figure 3b), as well as a fully connected graph $\mathcal{G}_{full}(\theta, T)$ over T frames (representing all possible sequence of letters, as in Figure 3c), ASG aims at minimizing:

$$ASG(\theta, T) = - \operatorname*{logadd}_{\pi \in \mathcal{G}_{asg}(\theta,T)} \sum_{t=1}^{T} (f_{\pi_t}(x) + g_{\pi_{t-1},\pi_t}(x)) + \operatorname*{logadd}_{\pi \in \mathcal{G}_{full}(\theta,T)} \sum_{t=1}^{T} (f_{\pi_t}(x) + g_{\pi_{t-1},\pi_t}(x)) \,, \tag{3}$$

where $g_{i,j}(\cdot)$ is a transition score model to jump from label i to label j. The left-hand part of 3 promotes sequences of letters leading to the right transcription, and the right-hand part demotes all sequences of letters. As for CTC, these two parts can be efficiently computed with the Forward algorithm. Derivatives with respect to $f_i(\cdot)$ and $g_{i,j}(\cdot)$ can be obtained (maths are a bit tedious) by applying the chain rule through the Forward recursion.

2.4 Beam-Search Decoder

We wrote our own one-pass decoder, which performs a simple beam-search with beam threholding, histogram pruning and language model smearing Steinbiss et al. (1994). We kept the decoder as

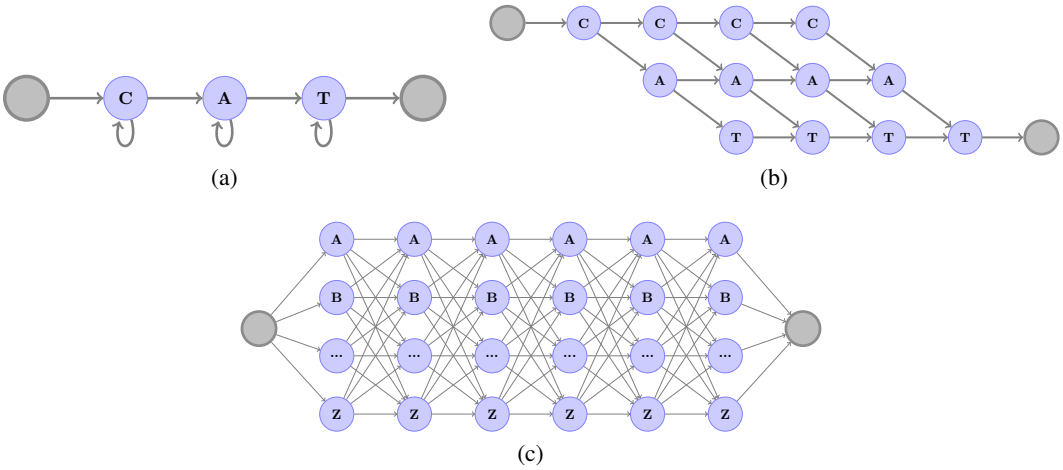


Figure 3: The ASG criterion graph. (a) Graph which represents all the acceptable sequences of letters for the transcription "cat". (b) Shows the same graph unfolded over 5 frames. (c) Shows the corresponding fully connected graph, which describe all possible sequences of letter; this graph is used for normalization purposes. Un-normalized transitions scores are possible on the edges. At each time step, nodes are assigned a conditional un-normalized score, output by the neural network acoustic model.

simple as possible (under 1000 lines of C code). We did not implement any sort of model adaptation before decoding, nor any word graph rescoring. Our decoder relies on KenLM Heafield et al. (2013) for the language modeling part. It also accepts un-normalized acoustic scores (transitions and emissions from the acoustic model) as input. The decoder attempts to maximize the following:

$$\mathcal{L}(\theta) = \underset{\pi \in \mathcal{G}_{asg}(\theta,T)}{\text{logadd}} \sum_{t=1}^{T} (f_{\pi_t}(x) + g_{\pi_{t-1},\pi_t}(x)) + \alpha \log P_{lm}(\theta) + \beta|\theta|\,, \tag{4}$$

where $P_{lm}(\theta)$ is the probability of the language model given a transcription θ, α and β are two hyper-parameters which control the weight of the language model and the word insertion penalty respectively.

3 Experiments

3.1 Setup

We implemented everything using Torch7[1]. The ASG criterion as well as the decoder were implemented in C (and then interfaced into Torch).

We consider as benchmark LibriSpeech, a large speech database freely available for download (Panayotov et al., 2015). LibriSpeech comes with its own train, validation and test sets. Except when specified, we used all the available data (about 1000h of audio files) for training and validating our models. We use the original 16 KHz sampling rate. The vocabulary $\mathcal{L}$ contains 30 graphemes: the standard English alphabet plus the apostrophe, silence, and two special "repetition" graphemes which encode the duplication (once or twice) of the previous letter (see Section 2.3).

The architecture hyper-parameters, as well the decoder ones were tuned using the validation set. In the following, we either report letter-error-rates (LERs) or word-error-rates (WERs). WERs have been obtained by using our own decoder (see Section 2.4), with the standard 4-gram language model provided with LibriSpeech[2].

[1] `http://www.torch.ch`.
[2] `http://www.openslr.org/11`.

Table 1: CTC vs ASG. CTC is Baidu's implementation. ASG is implemented on CPU (C with OpenMP). Timings (in *ms*) for small sequences (input frames: 150, letter vocabulary size: 28, transcription size: 40) and long sequences (input frames: 700, letter vocabulary size: 28, transcription size: 200) are reported in (a) and (b) respectively. (c) reports performance in LER. Timings include both forward and backward passes. CPU implementations use 8 threads.

(a)

batch size	CTC CPU	CTC GPU	ASG CPU
1	1.9	5.9	2.5
4	2.0	6.0	2.8
8	2.0	6.1	2.8

(b)

batch size	CTC CPU	CTC GPU	ASG CPU
1	40.9	97.9	16.0
4	41.6	99.6	17.7
8	41.7	100.3	19.2

(c)

	ASG	CTC
dev-clean	10.4	10.7
test-clean	10.1	10.5

MFCC features are computed with 13 coefficients, a 25 ms sliding window and 10 ms stride. We included first and second order derivatives. Power spectrum features are computed with a 25 ms window, 10 ms stride, and have 257 components. All features are normalized (mean 0, std 1) per input sequence.

3.2 RESULTS

Table 1 reports a comparison between CTC and ASG, in terms of LER and speed. Our ASG criterion is implemented in C (CPU only), leveraging SSE instructions when possible. Our batching is done with an OpenMP parallel for. We picked the CTC criterion implementation provided by Baidu[3]. Both criteria lead to the same LER. For comparing the speed, we report performance for sequence sizes as reported initially by Baidu, but also for longer sequence sizes, which corresponds to our average use case. ASG appears faster on long sequences, even though it is running on CPU only. Baidu's GPU CTC implementation seems more aimed at larger vocabularies (e.g. 5000 Chinese characters).

We also investigated the impact of the training size on the dataset, as well as the effect of a simple data augmentation procedure, where shifts were introduced in the input frames, as well as stretching. For that purpose, we tuned the size of our architectures (given a particular size of the dataset), to avoid over-fitting. Figure 4a shows the augmentation helps for small training set size. However, with enough training data, the effect of data augmentation vanishes, and both type of features appear to perform similarly. Figure 4b reports the WER with respect to the available training data size. We observe that we compare very well against Deep Speech 1 & 2 which were trained with much more data Hannun et al. (2014); Amodei et al. (2015).

Finally, we report in Table 2 the best results of our system so far, trained on 1000h of speech, for each type of features. The overall stride of architectures is 320 (see Figure 1), which produces a label every 20 ms. We found that one could squeeze out about 1% in performance by refining the precision of the output. This is efficiently achieved by shifting the input sequence, and feeding it to the network several times. Results in Table 2 were obtained by a single extra shift of 10 ms. Both power spectrum and raw features are performing slightly worse than MFCCs. One could expect, however, that with enough data (see Figure 4) the gap would vanish.

[3] `https://github.com/baidu-research/warp-ctc`.

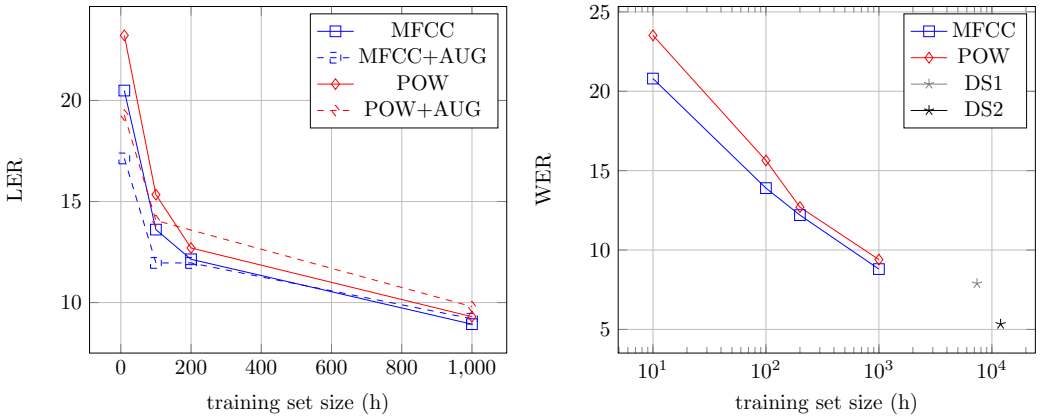


Figure 4: Valid LER (a) and WER (b) v.s. training set size (10h, 100h, 200h, 1000h). This compares MFCC-based and power spectrum-based (POW) architectures. AUG experiments include data augmentation. In (b) we provide Baidu Deep Speech 1 and 2 numbers on LibriSpeech, as a comparison Hannun et al. (2014); Amodei et al. (2015).

Table 2: LER/WER of the best sets of hyper-parameters for each feature types.

	MFCC		PS		Raw	
	LER	WER	LER	WER	LER	WER
dev-clean	6.9		9.3		10.3	
test-clean	6.9	7.2	9.1	9.4	10.6	10.1

4 Conclusion

We have introduced a simple end-to-end automatic speech recognition system, which combines a standard 1D convolutional neural network, a sequence criterion which can infer the segmentation, and a simple beam-search decoder. The decoding results are competitive on the LibriSpeech corpus with MFCC features (7.2% WER), and promising with power spectrum and raw speech (9.4% WER and 10.1% WER respectively). We showed that our AutoSegCriterion can be faster than CTC (Graves et al., 2006), and as accurate (table 1). Our approach breaks free from HMM/GMM pre-training and force-alignment, as well as not being as computationally intensive as RNN-based approaches (Amodei et al., 2015) (on average, one LibriSpeech sentence is processed in less than 60ms by our ConvNet, and the decoder runs at 8.6x on a single thread).

References

Dario Amodei, Rishita Anubhai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Jingdong Chen, Mike Chrzanowski, Adam Coates, Greg Diamos, et al. Deep speech 2: End-to-end speech recognition in english and mandarin. *arXiv preprint arXiv:1512.02595*, 2015.

L. R. Bahl, P. F. Brown, P. V. de Souza, and R. L. Mercer. Maximum mutual information estimation of hidden markov model parameters for speech recognition. In *Acoustics, Speech and Signal Processing (ICASSP), 1986 IEEE International Conference on*, pp. 49–52. IEEE, 1986.

Leon Bottou. *Une approche theorique de l'apprentissage connexionniste et applications a la reconnaissance de la parole*. PhD thesis, 1991.

Léon Bottou, Yoshua Bengio, and Yann Le Cun. Global training of document processing systems using graph transformer networks. In *Computer Vision and Pattern Recognition, 1997. Proceedings., 1997 IEEE Computer Society Conference on*, pp. 489–494. IEEE, 1997.

M. Gibson and T. Hain. Hypothesis spaces for minimum bayes risk training in large vocabulary speech recognition. In *Proceedings of INTERSPEECH*, pp. 2406—-2409. IEEE, 2006.

Alan Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. Speech recognition with deep recurrent neural networks. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pp. 6645–6649. IEEE, 2013.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, pp. 369–376. ACM, 2006.

Awni Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, et al. Deep speech: Scaling up end-to-end speech recognition. *arXiv preprint arXiv:1412.5567*, 2014.

Kenneth Heafield, Ivan Pouzyrevsky, Jonathan H Clark, and Philipp Koehn. Scalable modified kneser-ney language model estimation. In *ACL (2)*, pp. 690–696, 2013.

Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *Signal Processing Magazine, IEEE*, 29(6):82–97, 2012.

J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Eighteenth International Conference on Machine Learning, ICML*, 2001.

Yann LeCun and Yoshua Bengio. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361(10):1995, 1995.

Yajie Miao, Mohammad Gowayyed, and Florian Metze. Eesen: End-to-end speech recognition using deep rnn models and wfst-based decoding. *arXiv preprint arXiv:1507.08240*, 2015.

Abdel-rahman Mohamed, George E Dahl, and Geoffrey Hinton. Acoustic modeling using deep belief networks. *Audio, Speech, and Language Processing, IEEE Transactions on*, 20(1):14–22, 2012.

Dimitri Palaz, Ronan Collobert, and Mathew Magimai Doss. Estimating phoneme class conditional probabilities from raw speech signal using convolutional neural networks. *arXiv preprint arXiv:1304.1018*, 2013.

Dimitri Palaz, Mathew Magimai-Doss, and Ronan Collobert. Joint phoneme segmentation inference and classification using crfs. In *Signal and Information Processing (GlobalSIP), 2014 IEEE Global Conference on*, pp. 587–591. IEEE, 2014.

Dimitri Palaz, Ronan Collobert, et al. Analysis of cnn-based speech recognition system using raw speech as input. In *Proceedings of Interspeech*, number EPFL-CONF-210029, 2015.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: an asr corpus based on public domain audio books. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, pp. 5206–5210. IEEE, 2015.

Vijayaditya Peddinti, Guoguo Chen, Vimal Manohar, Tom Ko, Daniel Povey, and Sanjeev Khudanpur. Jhu aspire system: Robust lvcsr with tdnns, i-vector adaptation, and rnn-lms. In *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop*, 2015a.

Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur. A time delay neural network architecture for efficient modeling of long temporal contexts. In *Proceedings of INTERSPEECH*, 2015b.

Tara N Sainath, Ron J Weiss, Andrew Senior, Kevin W Wilson, and Oriol Vinyals. Learning the speech front-end with raw waveform cldnns. In *Proc. Interspeech*, 2015.

George Saon, Hagen Soltau, David Nahamoo, and Michael Picheny. Speaker adaptation of neural network acoustic models using i-vectors. In *ASRU*, pp. 55–59, 2013.

George Saon, Hong-Kwang J Kuo, Steven Rennie, and Michael Picheny. The ibm 2015 english conversational telephone speech recognition system. *arXiv preprint arXiv:1505.05899*, 2015.

Andrew Senior, Georg Heigold, Michiel Bacchiani, and Hank Liao. Gmm-free dnn training. In *Proceedings of ICASSP*, pp. 5639–5643, 2014.

Tom Sercu, Christian Puhrsch, Brian Kingsbury, and Yann LeCun. Very deep multilingual convolutional neural networks for lvcsr. *arXiv preprint arXiv:1509.08967*, 2015.

Hagen Soltau, George Saon, and Tara N Sainath. Joint training of convolutional and non-convolutional neural networks. In *ICASSP*, pp. 5572–5576, 2014.

Volker Steinbiss, Bach-Hiep Tran, and Hermann Ney. Improvements in beam search. In *ICSLP*, volume 94, pp. 2143–2146, 1994.

Philip C Woodland and Steve J Young. The htk tied-state continuous speech recogniser. In *Eurospeech*, 1993.