INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

CTC Training of Multi-Phone Acoustic Models for Speech Recognition

Olivier Siohan

Google, USA
siohan@google.com

Abstract

Phone-sized acoustic units such as triphones cannot properly capture the long-term co-articulation effects that occur in spontaneous speech. For that reason, it is interesting to construct acoustic units covering a longer time-span such as syllables or words. Unfortunately, the frequency distribution of those units is such that a few high frequency units account for most of the tokens, while many units rarely occur. As a result, those units suffer from data sparsity and can be difficult to train. In this paper we propose a scalable data-driven approach to construct a set of salient units made of sequences of phones called M-phones. We illustrate that since the decomposition of a word sequence into a sequence of M-phones is ambiguous, those units are well suited to be used with a connectionist temporal classification (CTC) approach which does not rely on an explicit frame-level segmentation of the word sequence into a sequence of acoustic units. Experiments are presented on a Voice Search task using 12,500 hours of training data.
Index Terms: acoustic modeling, CTC, multi-phone units, pronunciation modeling

1. Introduction

The performance of automatic speech recognition systems (ASR) has improved tremendously in the past several years due to the availability of very large amount of acoustic training data and the use of deep learning applied to both acoustic and language modeling [1]. A common aspect of those systems is that the acoustic model typically relies on the assumption inherited from traditional phonology that the speech signal can be decomposed as a sequence of atomic non-overlapping units [2]. The "beads-on-a-string" view of the speech signal leads to the use of sub-word units such as triphones which satisfy some of the desirable properties of good acoustic units. Specifically, as mentioned in [3] the acoustic units should be *trainable*, i.e. occur frequently enough in the training corpus to be estimated in a robust manner, *generalizable*, i.e. enable the modeling of words unseen in the training corpus, and *invariant* to the phonetic context to provide robustness to co-articulation effects and related factors.

Because co-articulation effects occur over a long time-span, units such as triphones may fail to capture some of the long-term temporal dependencies present in the speech signal. As a result, there has been many attempts to use longer acoustic units, either derived from lexical knowledge such as whole-word [4] or syllable units [5, 6, 7, 8, 9] , or automatically constructed from acoustic information [10, 11, 12, 13].

Since ASR systems are designed to minimize the word error rate, it may be attractive to select words as the basic acoustic unit. Unfortunately, the frequency distribution of words in natural language approximately follows a Zipf's law: most of the tokens in text are accounted for by a small number of high frequency words (e.g. function words such as "a", "the", "of") while there are many low frequency words. As a result, whole-word models are of a limited practical interest for large vocabulary ASR unless a very large amount of training data is available, such as in [4] where 125,000 hours of training data was used to enable the use of a large inventory of word units.

Syllable units received a lot of attention after studies on the Switchboard corpus illustrated that deviations from the canonical dictionary pronunciation are dependent on the syllable structure [14]. Despite the many attempts to build syllable-based systems [5, 6, 7, 8, 9], they never supplanted triphone-based systems, in part because many syllables lack coverage, hence requiring the development of hybrid syllable and phone-based systems, as well as due to the complexity in developing a high quality syllabification system to construct a pronunciation dictionary.

An attractive approach for defining acoustic units is to rely on the acoustic signal itself to identify homogeneous sound segments and cluster them into a set of units [10], enabling the joint design of the lexicon and the acoustic units [12, 13]. Unfortunately, those approaches do not properly scale as it is difficult to infer pronunciation for words not seen in training.

In this paper, we propose to use sequences of multiple phones called M-phones as acoustic units. By design, those units are constructed from a traditional phone-based pronunciation lexicon. An automatic procedure is used to construct the unit inventory and guarantees that the units are trainable and can model words unseen in the training corpus. The use of long sequences captures the long term temporal dependencies that are lacking in triphones. While M-phones have been proposed in [15] and [16], their evaluation focused on Gaussian mixture-based systems on small or medium vocabulary tasks such as TIMIT and Wall Street Journal using short phone sequences (e.g. phone-pairs in [15]), or on tasks consisting of numbers and isolated words recognition [16].

In contrast, we introduce in Section 2 a scalable and data-driven approach to construct an inventory of M-phone units consisting of the most salient sequences of phones. We describe a procedure to map any sequence of words into a graph-like structure of M-phones. In Section 3, we propose to use CTC [17, 18] to alleviate the need to explicitly decompose word sequences into M-phone sequences. The performance of the proposed approach is evaluated in Section 4 on a large Voice Search task using 12,500 hours of training data and 6 evaluation sets of 11 hours each. Section 5 concludes the paper.

2. Construction of the M-phone acoustic unit inventory

We define a multi-phone of order M, or M-phone for short, as a sequence of up to M phones. For example, assuming a system with a 2 phone inventory, a and b, the full set of M-phones of order 2 is defined as $\{a, b, a\_a, a\_b, b\_a, b\_b\}$. Since the total number of M-phones increases exponentially with the order M, one should define a procedure to limit the size of the M-phones inventory to the most salient sequences of phones. In [16], an iterative procedure is defined to grow a set of M-phones, where starting from a set of single phones, a pair of units u_1, u_2 is merged

http://dx.doi.org/10.21437/Interspeech.2017-505

into a new unit $u_1\_u_2$ based on the resulting increase in mutual information. The procedure guarantees that all single phones are kept in the final unit inventory, while the inventory to grow to an arbitrary size.

Similarly, we propose an approach which can scale to construct a set of M-phones of arbitrary size while enforcing that single phones are part of the inventory. This guarantees that the set of M-phone units generalizes to any word, as long as a grapheme-to-phoneme pronunciation engine is available. In our approach, we simply convert the word sequences representing the utterance transcripts from a large acoustic training set into their corresponding sequences of phones by running forced-alignment using an existing acoustic model. Next, we train a phone-based N-gram language model of order M on the phone sequences. Note that in all our experiments, we used Good-Turing smoothing when estimating the phone N-gram LM. The resulting set of phone N-grams directly defines the M-phone inventory. For a given order M, the size of the M-phone inventory can be controlled by pruning the LM using an entropy-based criterion [19]. Table 1 lists a few high-order M-phones and illustrates that M-phones extracts salient phone sequences, possibly spanning across word boundaries.

Table 1: *Example of M-phones acoustic units.*

M-phone	Matching word segment
@‘-dZ-I-n-j	V(irgini)a
O-r@-d-@	Fl(orida)
aU-m-E-n-i	H(ow Many)
d-r{-f-t	draft
f-O-rn-i	Cali(forni)a
g-u-g-@-l	Google

3. Training of the M-phone acoustic units

3.1. Mapping word sequences into M-phones sequences

Acoustic model training typically involves converting the word transcript of each training utterance into its corresponding sequence of acoustic units. With triphone-based acoustic models, this operation is often carried out via finite-state transducer (FST) composition [20].

Assuming that the following resources are available, a lexicon FST L mapping words into sequences of phones, a context-dependency transducer C mapping sequence of phones into sequences of HMMs, and an HMM transducer H mapping HMM sequences into context-dependent state sequences, then the reference transcript of a training utterance can be represented by a word-level acceptor T and turned into its corresponding sequence of CD-states by the FST composition $H \circ C \circ L \circ T$. This typically results in a linear or near-linear sequence of CD-states when accounting for pronunciation variants introduced by the lexicon and optional inter-word silence, which is suitable to enable the use of Viterbi training to estimate the CD-state models.

When using M-phone models, an additional issue is that there are multiple ways to decompose a word into its corresponding sequences of M-phones. For example, the word “mall”, phonetically transcribed as $/m//O//l/$ can be decomposed into M-phones as $/M-O-l/$, $/m//O-l/$, $/m-O//l/$, $/m//O//l/$. We propose to handle this conversion by constructing a transducer $\tilde{C}$, similar in role to the C transducer using in triphone-based system, designed to map sequences of phones into their corresponding graph of M-phones. For each M-phone in the M-phone inventory, a sequence of states is constructed, with the phone on the output label side and the M-phone on the input. An example of such a transducer is represented in Fig. 1 showing how the $/m-O-l/$ and $/m-O/$ M-phones are constructed.

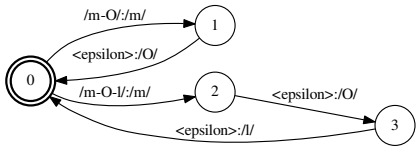

Figure 1: *Illustration of the construction of an M-phone $\tilde{C}$ transducer with only 2 M-phones. The transducer converts a $/m//O/$ phone sequence into $/m-O/$ and $/m//O//l/$ into $/m-O-l/$.*

Note that since the M-phone units are constructed using a C-like transducer which operates on sequences of phones, M-phones are not constrained to be within-words. Instead, an M-phone unit can span across-words, consistent with the way the M-phone inventory was designed from a phone-based LM. This is illustrated in Fig. 2 which represents the M-phone pronunciation network for the sentence “promenade mall”. The $/d-m/$ and $/d-m-O/$ M-phones are examples of M-phone crossing word boundaries. Note also that the construction of the M-phone $\tilde{C}$ transducer implies that the M-phone models are context-independent. Instead, long-term temporal dependencies are implicitly modeled via the use of long M-phone units.

3.2. CTC Training

The representation of a word sequence as a graph-like structure of M-phones as in Fig. 2 suggests that M-phones acoustic units are not amenable to Viterbi-style training, especially since the parallel path structure consists of acoustic units competing with many similar sub-units.

However, the CTC training procedure introduced in [17] and applied to speech recognition in [18] has been shown to automatically learn a correspondence between a sequence of input frames and its corresponding sequence of output labels, without the need for explicitly constructing a frame-level segmentation of the output labels. This is achieved by using a special output label called *blank* used to model the probability of not emitting any output label for a given input frame, thus enabling the input and output sequences to have different lengths. In CTC, a target label sequence such as “$a\ b\ c$” defines an equivalence class represented by the regular expression “$blank^*\ a^+\ blank^*\ b^+\ blank^*\ c^+\ blank^*$”. This enables the output label sequence to “stretch” as needed to match the length of the input sequence. More formally, let x represent a sequence of input frames of length T, y a sequence of output labels of length N with $N \leq T$, and $\Phi(y)$ a set of sequences of labels of length T constructed from the equivalence label sequence class of y. The CTC training optimizes the model parameter to maximize the likelihood of the output label sequence given the input sequence, defined as:

$$P(y|x) = \sum_{p \in \Phi(y)} P(p|x). \quad (1)$$

The summation over all the possible paths p encoding the alignments of the output label sequence with the input sequence can be done with the forward-backward algorithm.

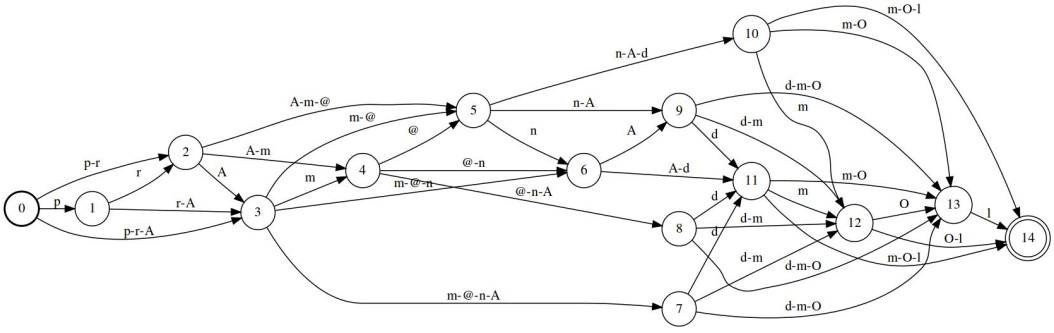

Figure 2: *Example of an M-phone pronunciation network for the word sequence "promenade mall".*

Since CTC training accommodates output sequences of variable length for a given input sequence without relying on an explicit frame-level segmentation, it becomes an attractive approach to train M-phone acoustic units as the forward-backward algorithm integrates over all possible decompositions of a word sequence into M-phones. A similar argument was recently used in [21] but using character N-grams rather than phone N-grams are basic acoustic units. In addition, it was shown in [4] that the use of an LSTM model combined with CTC enables the use of a single-state HMM topology to represent whole words. This confirms that LSTM combined with a CTC loss function can learn long acoustic units spanning a large temporal window.

4. Experiments and Results

4.1. Database

All experiments are carried out using sets of English queries representative of Google's voice search traffic that were anonymized and hand-transcribed. Those queries are about 3 sec. long on average. Acoustic models are trained using a multi-style training procedure [22] where the original queries are artificially corrupted using a room simulator to add varying degrees of noise and reverberation and simulate a far-field environment. The noise sources consist of music sampled from YouTube as well as recordings of daily life environments.

The training set is constructed from a total of 15 million utterances (about 12,500 hours) artificially corrupted using the room simulator configured to generate utterances with a signal-to-noise ratio (SNR) ranging from 0 to 30 dB with a 12 dB average from millions of room configurations with a T60 reverberation time ranging from 0 to 900 ms for an average of 480 ms and an average source-to-microphone distance of 3 m. This multi-style training set is designed to improve the robustness of our system to noise.

An evaluation set of 13k Google voice search utterances called *clean* is used to construct 5 additional sets by perturbing the data in various ways. The *noisy1* set consists of utterances generated using the room simulator with an average SNR of 16 dB, a T60 of 170 ms and a source-to-microphone distance of 0.7 m. The *noisy3* set was also constructed using the room simulator but configured to match the training acoustic conditions. The *rerecorded* set is constructed by re-recording the original *clean* set played using a mouth simulator from 5 different locations in a living room environment, at distances ranging from 1.6 m to 4.4 m from the microphone. Last, both the *rerecorded noisy* and *rerecorded multi* sets are constructed by adding noise to the *rerecorded* set, restricted to YouTube music-like noises for *rerecorded noisy* and YouTube speech-like noises for *rerecorded multi*, with an SNR between 0 dB and 20 dB but biased towards 5 dB and 10 dB.

4.2. Results

All experiments are conducted using a unidirectional LSTM model consisting of 5 LSTM layers of 600 units each. The input features consist of 128 log-mel energy coefficient computed every 30 ms. The order of the M-phone model was set to 5 and the size of the acoustic unit inventory was set to 1667 including the blank symbol. The acoustic unit inventory did not include any silence unit as we found it unnecessary for CTC training. We did not use any special markers to represent word start/end. Some preliminary experiments suggested that performance was not affected much by the size of the acoustic unit inventory provided that it was larger than 1.5k units.

All training experiments were carried out starting from a randomly initialized model. The training did not use any presegmentation information, which is sometimes used to constrain the range of valid alignments during CTC training [18].

Given that a short word sequence such as the one represented in Fig. 2 can be decomposed into many sequences of M-phones, it is interesting to observe examples of decompositions that the training selects. In Table 2, we extracted a random snippet from our training logs representing the sequence of M-phone labels predicted by the model along training for a few training utterances. For each utterance, REF represents the reference word-level transcript while HYP represent the sequence of M-phone labels predicted by the network. Note that HYP is constructed by scoring each input frame, selecting the highest scoring output symbol (either *blank* or one of the M-phones) and discarding repetitions and blanks. For example, the predicted frame-level sequence $a\ a\ a\ blank\ b\ b\ blank$ is transformed into the output label sequence $a\ b$.

One can observe that the predicted output label sequences mostly contain long M-phones and the model rarely predicts a sequence of individual phones. This can be explained by the fact that CTC training leads to an acoustic model typically outputting a single spike for each output label and blank otherwise, as was demonstrated in many studies on CTC (e.g. [18, 4]). Given that predicting blank is a high likelihood event, when given a graph structure like the one in Fig. 2, CTC will prefer paths reaching the final state in the shortest number of hops. This favors paths with the longest M-phones because those paths emit the largest possible number of *blank* symbols. This fits with our objective of decomposing the speech signal into long acoustic units, without any explicit need of a duration model.

Table 2 also illustrates some anecdotal evidence from our logs that suggests a syllable-like structure of the predicted M-phones that delineates individual words. For example, in the last utterance, each individual word is predicted as a single M-phone unit.

This is interesting since the entire procedure for constructing the M-phone inventory and training the acoustic model did not rely on any explicit word boundary information. The units were defined from sequences of phones and no constraints were used to force the training to prefer decomposing a graph of M-phones into its longest units. This suggests that both the procedure used to define the M-phone unit inventory by constructing a phone LM and the CTC training procedure succeeded in extracting long salient acoustic units in a fully data-driven manner.

Table 2: *Example of predicted M-phone sequences during training, along with their corresponding reference word-level transcript.*

HYP: /S-oU/ /m-i/ /{/ /p-I-k/ /tS-@‘/ /V-v/
REF: Show me a picture of
HYP: /w-V-t/ /d-V-z/ /p-r-@/ /p-O-r/ /S-@-n/ /@-l/ /t-u/ /m-i-n/
REF: What does proportional to mean?
HYP: /j-V-N/ /s-@‘/ /s-aU-n-d/ /k-l-aU-d/
REF: John Se SoundCloud
HYP: /w-E-r/ /{/ /m-aI/
REF: Where am I?
HYP: /d-@/ /m-i/ /n-@‘/ /s-p-E-l/ /@-N/
REF: demeanor spelling
HYP: /w-E-r/ /k-{-n/ /aI/ /g-E-t/ /g-{-s/
REF: Where can I get gas?

We compared a system constructed on the proposed M-phone units again a competitive baseline constructed using a set of 9k CD-phones as acoustic units. Experimental results are given on 6 evaluation sets in terms of word error rate (WER) in Table 3. While the M-phone based system did not outperform the well tuned baseline, the difference in performance were less than 7.3% relative across the different test sets.

Table 3: *Word Error Rates of an M-phone system compared to a CD-phones baseline on multiple test sets.*

Eval Set	CD-phones	M-phones	WER Reduction
clean	12.7	13.1	-3.1%
noisy1	15.0	16.1	-7.3%
noisy3	20.1	21.5	-7.0%
rerecorded	22.7	23.6	-4.0%
rerecorded_multi	45.8	48.4	-5.7%
rerecorded_noisy	34.6	36.3	-4.9%

One hypothesis to explain the performance of the M-phone system was that the number of possible decomposition of a word sequence into M-phones is too large and should be constrained. We experimented with an approach to prefer long M-phones over short ones by constructing an M-phone to phone C transducer weighting each M-phone with a cost inversely proportional to its length. This did not result in improved performance, in part because the weighting still led to many paths with identical costs and mostly discarded the most unlikely single-phone based decomposition.

Another concern was related to the discrepancy in computing the score of a sequence of words between training and test. At training time, the forward-backward algorithm is used and considers the contribution of all possible decompositions of a sequence of words into M-phones. In contrast, at test time, the recognizer computes the score of each hypothesis using the Viterbi approximation. Referring again to Fig. 2, this means that at training time, the likelihood of the corresponding word sequence is computed over all possible paths, while during recognition only the score of the best path is used. For that reason, we implemented an N-best rescoring approach to rescore each word sequence hypothesis using its forward score, by constructing a rescoring network similar to the one in Fig. 2. The acoustic score of each hypothesis in the N-best list was recomputed and combined with the original LM score and the resulting total score was used to rank the N-best list. We first evaluated the N-best oracle WER both on a competitive production model as well as on our M-phone based system, and confirmed that in both cases, the 5-best oracle WER was about 40% lower than the 1-best WER, with the 2-best oracle WER being 20% lower than the 1-best. This confirmed that an N-best rescoring procedure could potentially improve performance. Rescoring results are presented in Table 4 but using an older system hence not directly comparable to the previous results. It was shown that the rescoring led to a slight degradation of the WER. Similar results were observed when rescoring the traditional CD-phone CTC system. Hence, the discrepancy in acoustic score computation alone does not explain that so far, the M-phone based system failed to outperform a well-turned CD-phone system.

Table 4: *Impact of N-best rescoring using the CTC forward score.*

Eval Set	1-best	5-best rescoring	50-best rescoring
clean	14.4	14.8	14.9
noisy1	16.9	17.3	17.4
noisy3	22.4	22.6	22.8
rerecorded	24.1	24.4	24.5
rerecorded_multi	48.7	49.2	49.5
rerecorded_noisy	36.2	36.5	36.7

5. Conclusions

We proposed a fully data-driven approach to construct a set of long acoustic units for the purpose of modeling long temporal dependencies in the speech signal. We showed that the use of sequences of phones of variable length called M-phones fits well with an LSTM model and a CTC training criterion, which does not require an explicit decomposition of a sequence of words into its corresponding sequence of M-phones. We illustrated that the approach discovered word-like units and that the CTC criterion naturally favors long units over shorter ones. As recently described in [21] in an approach very similar to ours but operating on sequence of characters, this presents the additional advantage of alleviating the conditional independence assumption of the output labels in CTC, by predicting clusters of phones as a single label. The proposed approach evaluated on a large task with over 12,500 hours of training data did not outperform a competitive CD-phone baseline but within a range of 3.1% to 7.3% relative WER difference depending on the eval set. We believe that the performance of the M-phone system could be improved by replacing the single phone units by a set of CD-phone units conditioned on the first and last phone of the preceding/following unit. This should prevent the sudden loss of contextual information when a single phone occurs surrounded in a sequence of M-phones.

6. References

[1] G. Hinton, L. Deng, D. Yu, G. Dahl, A. R. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[2] A. H. Sommerstein, *Modern Phonology*. University Park Press, 1977.

[3] K. Livescu, E. Fosler-Lussier, and F. Metze, "Subword modeling for automatic speech recognition: Past, present, and emerging approaches," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 44–57, Nov. 2012.

[4] H. Soltau, H. Liao, and H. Sak, "Neural speech recognizer: acoustic-to-word LSTM model for large vocabulary speech recognition," Oct. 2016, arXiv:1610.09975.

[5] R. J. Jones, S. Downey, and J. S. Mason, "Continuous speech recognition using syllables," in *European Conference on Speech Communication and Technology (EuroSpeech)*, Rhodes, Greece, Sep. 1997, pp. 1171–1174.

[6] A. Hämäläinen, L. Boves, and J. de Veth, "Syllable-length acoustic units in large-vocabulary continuous speech recognition," in *Proceedings of SPECOM*, 2005, pp. 499–502.

[7] A. Hämäläinen, L. Boves, J. de Veth, and L. t. Bosch, "On the utility of syllable-based acoustic models for pronunciation variation modelling," *EURASIP Journal Audio Speech Music Process.*, vol. 2007, no. 2, pp. 3–3, Apr. 2007.

[8] A. Ganapathiraju, J. Hamaker, J. Picone, M. Ordowski, , and G. R. Doddington, "Syllable-based large vocabulary continuous speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 4, pp. 358–366, 2001.

[9] A. Sethy, B. Ramabhadran, , and S. Narayanan, "Improvements in English ASR for the MALACH project using syllable-centric models," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2003, pp. 129–134.

[10] R. C. Rose and E. Lleida, "Speech recognition using automatically derived acoustic baseforms," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 2, Apr. 1997, pp. 1271–1274.

[11] M. Ostendorf, "Moving beyond the "beads-on-a-string" model of speech," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 1999, pp. 79–84.

[12] M. Bacchiani and M. Ostendorf, "Using automatically-derived acoustic sub-word units in large vocabulary speech recognition," in *International Conference on Speech and Language Processing (ICSLP)*, Australia, 1998.

[13] R. Singh, B. Raj, and R. M. Stern, "Automatic generation of subword units for speech recognition systems," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 2, pp. 89–99, 2002.

[14] S. Greenberg, "Speaking in shorthand - a syllable-centric perspective for understanding pronunciation variation," *Speech Communication*, vol. 29, no. 2, pp. 159–176, 1999.

[15] P. O'Neill, S. Vaseghi, B. Doherty, W.-H. Tan, and P. M. McCourt, "Multi-phone strings as subword units for speech recognition," in *International Conference on Speech and Language Processing (ICSLP)*, 1998.

[16] R. Messina and D. Jouvet, "Context dependent "long units" for speech recognition," in *International Conference on Speech and Language Processing (ICSLP)*, Jeju Island, Korea, Oct. 2004.

[17] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *International Conference on Machine Learning (ICML)*, 2006, pp. 369–376.

[18] H. Sak, A. W. Senior, K. Rao, O. Irsoy, A. Graves, F. Beaufays, and J. Schalkwyk, "Learning acoustic frame labeling for speech recognition with recurrent neural networks," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 4280–4284.

[19] A. Stolcke, "Entropy-based pruning of backoff language models," in *Proceedings DARPA Broadcast News Transcription and Understanding Workshop*, 1998, pp. 270–274.

[20] M. Mohri, F. Pereira, and M. Riley, "Weighted finite-state transducers in speech recognition," *Computer Speech and Language*, vol. 16, pp. 69–88, 2002.

[21] H. Liu, Z. Zhu, X. Li, and S. Satheesh, "Gram-CTC: Automatic unit selection and target decomposition for sequence labelling," Mar. 2017, arXiv:1703.00096.

[22] R. P. Lippmann, E. A. Martin, and D. B. Paul, "Multi-style training for robust isolated word speech recognition," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1987.