WHEN CTC TRAINING MEETS ACOUSTIC LANDMARKS

Di He[*,1,2], *Xuesong Yang*[*,1,3], *Boon Pang Lim*[2], *Yi Liang*[1], *Mark Hasegawa-Johnson*[1], *Deming Chen*[1]

[1]ECE, Coordinated Science Lab, and Beckman Institute, University of Illinois, Urbana, IL, USA
[2]Novumind Inc, Santa Clara, CA, USA [3]Amazon Alexa Speech, Seattle, WA, USA

ABSTRACT

Connectionist temporal classification (CTC) provides an end-to-end acoustic model (AM) training strategy. CTC learns accurate AMs without time-aligned phonetic transcription, but sometimes fails to converge, especially in resource-constrained scenarios. In this paper, the convergence properties of CTC are improved by incorporating acoustic landmarks. We tailored a new set of acoustic landmarks to help CTC training converge more rapidly and smoothly while also reducing recognition error rates. We leveraged new target label sequences mixed with both phone and manner changes to guide CTC training. Experiments on TIMIT demonstrated that CTC based acoustic models converge significantly faster and smoother when they are augmented by acoustic landmarks. The models pretrained with mixed target labels can be further finetuned, resulting in phone error rates 8.72% below baseline on TIMIT. Consistent performance gain is also observed on WSJ (a larger corpus) and reduced TIMIT (smaller). With WSJ, we are the first to succeed in verifying the effectiveness of acoustic landmark theory on a mid-sized ASR task.

Index Terms— Acoustic Modeling, CTC, Acoustic Landmarks, End-to-End

1. INTRODUCTION

Automatic speech recognition (ASR) is a sequence labeling problem that translates a speech waveform into a sequence of words. Recent success of hidden Markov model (HMM) combined with deep neural networks (DNNs) or recurrent neural networks has achieved a word error rate (WER) on par with human transcribers [1, 2]. These hybrid acoustic models (AMs) are typically optimized by cross-entropy (CE) training which relies on accurate frame-wise context-dependent state alignments pre-generated from a seed AM. The connectionist temporal classification (CTC) loss function [3], in contrast, provides an alternative method of AM training in an end-to-end fashion—it directly addresses the sequence labeling problem without prior frame-wise alignments. CTC is capable of learning to construct frame-wise paths implicitly bridging between the input speech waveform and its context-independent target, and it has been demonstrated to outperform hybrid HMM systems when the amount of training data is large [4, 5, 6]. However, its performance degrades and is even worse than traditional CE training when applied to small-scale data [7].

Training CTC models can be time-consuming and sometimes models are apt to converge to even a sub-optimal alignment, especially on resource-constrained data. In order to alleviate such common problems of CTC training, additional tricks are needed, for example, ordering training utterances by their lengths [6] or bootstrapping CTC models with models CE-trained on fixed alignments [8]. The success of bootstrapping with prior alignments indicates that external phonetic knowledge may help to regularize CTC training towards stable and fast convergence. Furthermore, another investigation [9] reveals that the spiky predictions of CTC models tend to overlap with the vicinity of acoustic landmarks where abrupt manner changes of articulation occur [10]. The possible coincidence of CTC peaks overlapping acoustic landmarks suggests a number of possible approaches for reducing the data requirements of CTC, including cross-language transfer (using the relative language-independence of acoustic landmarks [11]) and informative priors.

Many efforts have been attempted to augment acoustic modeling with acoustic landmarks [11, 12, 13] which are detected by accurate time-aligned phonetic transcriptions. To the best of our knowledge, only TIMIT [14] (5.4 hours) provides such fine-grained transcriptions. The value of testing these approaches are limited since the only available corpus is very small. It is worth further exploring the power of landmark theory when scaled up to large corpus speech recognition.

In this paper, we propose to augment phone sequences with acoustic landmarks for CTC acoustic modeling and leverage a two-phase training procedure with pretraining and finetuning to address CTC convergence problems. Experiments on TIMIT demonstrate that our approaches not only help CTC models converge more rapidly and smoothly, but also achieve a lower phone error rate, up to 8.72% phone error rate reduction over CTC baseline with phone labels only. We also investigate the sensitivity of our approaches to the size of training data on subsets of TIMIT (smaller corpora) and

[*]Authors contributed equally.

WSJ [15] (a larger corpus). Our findings demonstrate that label augmentation generalizes to larger and smaller training datasets, and we believe this is the first work that applies acoustic landmark theory to a mid-sized ASR corpus.

2. BACKGROUND

2.1. Connectionist Temporal Classification (CTC)

Recent end-to-end systems have attracted much attention, for example, because they avoid time-consuming iterations between alignment and model building [3, 16]. The CTC loss computes the total likelihood of the target label sequence over all possible alignments given an input feature sequence, so that the computation is more expensive than frame-wise cross-entropy training. A blank symbol is introduced to compensate for the difference in length between an input feature sequence and its target label sequence. Forward-backward algorithms are used to efficiently sum the likelihood over all possible alignments. The CTC loss is defined as,

$$\mathcal{L}_{ctc} = -\log p\left(\mathbf{y}|\mathbf{x}\right) = -\log \sum_{\boldsymbol{\pi} \in \mathcal{B}^{-1}(\mathbf{y})} p(\boldsymbol{\pi}|\mathbf{x})$$

where $\mathbf{x}$ is an input feature sequence, $\mathbf{y}$ is the target label sequence of $\mathbf{x}$, $\boldsymbol{\pi}$ is one of blank-augmented alignments of $\mathbf{y}$, and $\mathcal{B}^{-1}(\mathbf{y})$ calculates the set of all such alignments. During decoding, the n-best list of predicted label sequences can be achieved by either a greedy search or a beam search based on weighted finite state transducers (WFSTs). In the following experiments, our acoustic models are trained by the phoneme CTC loss, and we report phone error rates on TIMIT (a smaller corpus) through an one-best greedy search and word error rates on WSJ (a larger corpus) through an one-best WFSTs beam search, respectively.

2.2. Acoustic Landmarks

Acoustic landmark theory originates from experimental studies of human speech production and speech perception. It claims there exist instantaneous acoustic events that are perceptually salient and sufficient to distinguish phonemes [10]. Automatic landmark detectors can be knowledge-based [17] or learned [18]. Landmark-based ASR has been shown to slightly reduce the WER of a large-vocabulary speech recognizer, but only in a rescoring paradigm using a very small test set [18]. Landmarks can reduce computational load for DNN/HMM hybrid models [12, 13] and can improve recognition accuracy [11]. Previous works [11, 12, 13, 19] annotated landmark positions mostly following experimental findings presented in [20, 21]. Four different landmarks are defined to capture positions of vowel peak, glide valley in glide-like consonants, oral closure and oral release.

3. METHODS

3.1. Distinctive Features and Landmark Definition

Distinctive features (DFs) concisely describe sounds of a language at a sub-segmental level, and they have direct relations to acoustics and articulation. These features take on binary encodings of perceptual, phonological, and articulatory speech sounds [22]. A collection of these binary features can distinguish each segment from all others in a language. Autosegmental phonology [23] also suggests that DFs have an internal organization with a hierarchical relationship with each other. We follow these linguistic rules to select two primary features—*sonorant* and *continuant*—that distinguish among the manner classes of articulation, resulting in a four-way categorization shown in Table 1. We define landmarks to be changes in the value of one of these two distinctive features using the TIMIT phone inventory. The standard phoneme set used by WSJ ignores detailed annotations of oral closures, for example /bcl/, so that we merge together [-,+*continuant*] features under [-*sonorant*] column in Table 1, resulting in a three-way categorization for WSJ experiments instead.

Table 1. Broad classes of sounds on TIMIT

Manner	-sonorant	+sonorant
-continuant	bcl dcl gcl kcl pcl q tcl	em en eng m n ng
+continuant	b d g k p t ch jh dh f hh hv s sh th v z zh	aa ae ah ao aw ax ax-h axr ay dx eh el ey ih ix iy l nv ow oy r uh uw ux w y er

3.2. Augmenting Phone Sequences With Landmarks

We defined two methods of augmenting phone label sequences with acoustic landmarks. *Mixed Label 1* only inserts landmarks between two broad classes of sounds where manner changes occur; *Mixed Label 2* inserts landmarks between phones even if manner changes don't exist. Figure 1 demonstrates an example of our two augmentation methods.

CTC only requires a single target label sequence, so that augmenting phone sequences with landmarks can relax the need for time-aligned phone transcriptions. With a blank label present between two phones in the training target sequence, the vanilla CTC training can be considered as already experimenting with the scenario where a dedicated phone boundary label is added to the label set. CTC is thus an ideal baseline for our experiments.

3.3. Acoustic Modeling using CTC

We follow a pretraining and finetuning procedure to train our CTC models. At the phase of pretraining, the AM initializes weights randomly and is trained by one of our mixed label

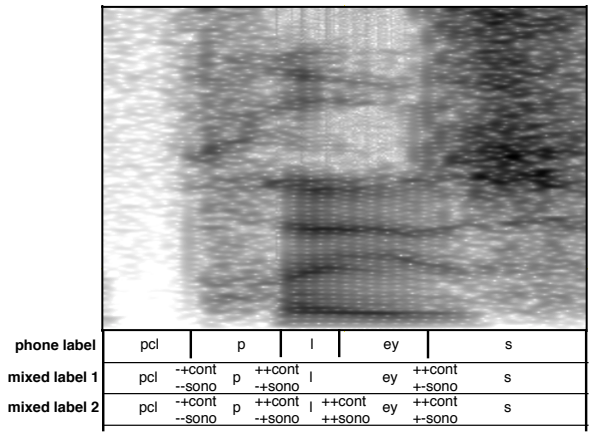

Fig. 1. Examples of target label sequences for the word "PLACE". The audio clip is selected from `SI792` on TIMIT.

sequences until convergence; at the phase of finetuning, the AM initializes weights from the pretrained model and continues to be trained by a label sequence with only phones. These two phases of training take the same acoustic features. Figure 2 briefly illustrates the whole procedure. The top output layer calculates a posterior distribution over symbols combined with both phones and landmarks, while the bottom output layer calculates it over only phones.

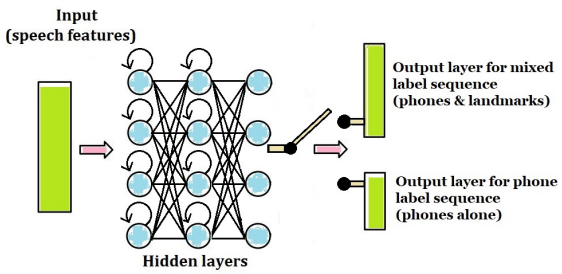

Fig. 2. Two-phase acoustic modeling: top output layer pretrains with mixed labels and bottom output layer finetunes with phone labels only

4. EXPERIMENTS

4.1. Configurations

We conducted our experiments on both the TIMIT [14] and WSJ [15] corpora. We used 40-dimensional log mel filterbank energy features computed with 10ms shift and 20ms span. No delta features or frame stacking were used. The recurrent neural networks stacked two layers of bidirectional LSTMs, each with 1024 cells (512 cells per direction), capped by a fully connected layer with 256 neurons. Weights are initialized randomly from Xavier uniform distribution [24]. New-Bob annealing [25] is used for early stopping after a minimum waiting period of two epochs. The initial learning rate is 0.0005. The TIMIT baseline is trained on 61 phones. The WSJ baseline is trained on 39 phones[1] defined in the CMU pronunciation dictionary. One-best greedy search is applied to calculate the phone error rate (PER). We did not map TIMIT phones to CMU phone set (39 phones). In order to make a fair comparison, all baselines went through the same two-phase training with pretraining and finetuning. One-best beam search based on WFSTs is applied to calculate the word error rate in WSJ experiments using decoding graphs with a primitive trigram (tg) and pruned trigram (tgpr) from EESEN[2]. We use the same train/dev/test split from Kaldi Recipes for TIMIT and WSJ.

4.2. Experiments on TIMIT

Figure 3 presents the development set PER as a function of training epoch. The PER for mixed sequence represented by the red and yellow lines in Figure 3 is calculated after landmark labels have been removed from the output sequence. In the pretrain phase, models trained on augmented labels do not seem to have any advantage in terms of error rate. However, the models converge much more rapidly and smoothly. After pretraining, both the baseline and mixed-label systems are finetuned; the mixed-label system (purple line in Fig. 3) returns a model that is more accurate.

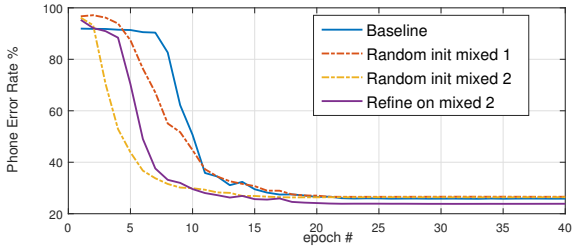

Fig. 3. PER as a function of training epoch. PER is calculated against only phones after landmarks are removed.

The exact PERs for different setups on the TIMIT test set are reported in Table 2. Our baseline achieved a PER of 30.36%, which was not improved by finetuning. This is higher than PER reported elsewhere (e.g., [3]), because nobody else calculates PER on the full TIMIT set of 61 phones. As shown in Table 2, if we train with mixed labels and strip away landmarks from the hypothesis sequence, landmarks provide little benefit. However, the *Mixed 1* and *Mixed 2* systems achieved lower PER after the finetuning stage by 4.64% and 8.72% relative, respectively. Apparently, a phone sequence augmented with landmarks can be learned more accurately than a raw phone sequence, perhaps because the acoustic features of manner transitions are easy to learn, and help to time-align the training corpus. The *Mixed Label 2* set outperforms *Mixed Label 1*, apparently because the ex-

[1] `https://github.com/Alexir/CMUdict/blob/master/cmudict-0.7b.phones`

[2] `https://github.com/srvk/eesen/blob/master/asr_egs/wsj/run_ctc_phn.sh`

tra boundary information in *Mixed Label 2* is useful to the training algorithm.

Table 2. Comparison between baseline and our proposed models with augmented target labels in PER (%). Number in the parentheses denotes the relative reduction over baseline.

	Baseline	Mixed 1	Mixed 2
random init	30.36	30.98	29.10
finetuned	30.36	**28.96** (4.64%)	**27.72** (8.72%)

It is not clear why a finetuning stage is needed in order for *Mixed 1* to beat the baseline. One possibility is that landmark labels are helpful for some tokens, and harmful for others; pretraining uses the helpful landmarks to learn better phone alignments, then finetuning permits the network to learn to ignore the harmful landmark tokens. We looked into the prior distribution on TIMIT, presented in Figure 4, of both phones (top subplot, with phones ordered in the same way as they occurred in Table 1) and landmarks (bottom subplot, *Mixed Label 2* ordered in category permutation using *continuant* as the first variable and *sonorant* as the second). The table reveals that the distribution of landmarks is not balanced. Most labels indicate a transition related to the [+*continuant*,+*sonorant*] phones. A skewed landmark support is not ideal for augmenting phone recognizer training as it tends to provide the same and redundant information for many training sequences.

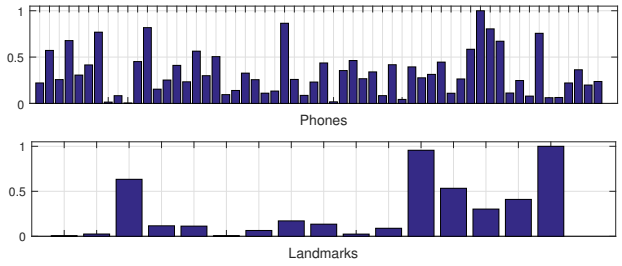

Fig. 4. Prior distributions of phones and acoustic landmarks.

4.3. Datasets Smaller and Larger than TIMIT

To solidify our findings, we further investigated the sensitivity of our approaches to the size of training data on subsets of TIMIT (smaller corpora) and WSJ (a larger corpus). In this section, we only demonstrate the experiments using *Mixed Label 2* augmentation method since it outperforms *Mixed Label 1* in the previous discussion. We report PER/WER results for finetuned models.

Figure 5 shows the PER results by stretching the amount of training data on TIMIT. Both the proposed model and baseline fail to converge when 75% of the training data is used. We observe that both models start to predict a constant sequence (usually made up of two to three most frequent phones) for all utterances. Scheduled reducing the learning rate by New-Bob annealing can't help to converge to an optimal. Increasing the amount of training data helps both models converge. The baseline needs 90% of TIMIT to converge, while the proposed system only needs 80% of TIMIT.

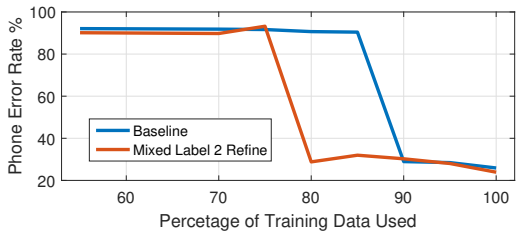

Fig. 5. PERs by stretching the amount of training data on TIMIT.

When scaling up to a even larger corpus on WSJ, the proposed *Mixed Label 2* system could achieve better performance over the baseline consistently in terms of all metrics as shown in Table 3. Our baseline system slightly underperforms the results published in EESEN [5] because our network is shallower and the acoustic inputs do not include any dynamic (delta) features, but the benefit of the proposed landmark augmentation method still applies. To our knowledge, this is the first work to show that manner-change acoustic landmarks reduce both PER and WER on a mid-sized ASR corpus.

Table 3. Label Error Rate (%) on WSJ, where tg and tgpr denote decoding graphs with primitive and pruned trigrams.

	PER		WER (tgpr / tg)	
	eval92	dev93	eval92	dev93
Baseline	8.7	12.38	8.75/8.17	13.15/12.31
Mixed 2	**8.12**	**11.49**	**8.35**/8.19	**12.86/12.28**

5. CONCLUSION

We proposed to augment CTC with acoustic landmarks. We modified the classic landmark definition to suit the CTC criterion and implemented a pretraining-finetuning training procedure to improve CTC AMs. Experiments on TIMIT and WSJ demonstrated that CTC training becomes more stable and rapid when phone label sequences are augmented by landmarks, and achieves a significantly lower (8.72% relative reduction) asymptotic PER. The advantage is consistent across corpora (TIMIT, WSJ) and across metrics (PER, WER). CTC with landmarks converges when the dataset is too small to train the baseline, and it also converges without the need of time alignments on a mid-sized standard ASR training corpus (WSJ).

Acknowledgements: The fifth author was supported by the DARPA LORELEI program. All results and conclusions are those of the authors, and are not endorsed by DARPA.

6. REFERENCES

[1] Wayne Xiong, Jasha Droppo, Xuedong Huang, Frank Seide, Mike Seltzer, Andreas Stolcke, Dong Yu, and Geoffrey Zweig, "Toward human parity in conversational speech recognition," *IEEE/ACM Trans. Audio Speech and Language*, vol. 25, no. 12, pp. 2410–2423, 2017.

[2] George Saon, Gakuto Kurata, Tom Sercu, Kartik Audhkhasi, Samuel Thomas, Dimitrios Dimitriadis, Xiaodong Cui, Bhuvana Ramabhadran, Michael Picheny, Lynn-Li Lim, et al., "English conversational telephone speech recognition by humans and machines," in *Interspeech*. ISCA, 2017, pp. 132–136.

[3] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *International Conference on Machine Learning*. ACM, 2006, pp. 369–376.

[4] Haşim Sak, Andrew Senior, Kanishka Rao, and Françoise Beaufays, "Fast and accurate recurrent neural network acoustic models for speech recognition," in *Interspeech*. ISCA, 2015, pp. 1468–1472.

[5] Yajie Miao, Mohammad Gowayyed, and Florian Metze, "EESEN: End-to-end speech recognition using deep RNN models and WFST-based decoding," in *ASRU*. IEEE, 2015, pp. 167–174.

[6] Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang Chen, et al., "Deep speech 2: End-to-end speech recognition in English and Mandarin," in *International Conference on Machine Learning*, 2016, pp. 173–182.

[7] Yajie Miao, Mohammad Gowayyed, Xingyu Na, Tom Ko, Florian Metze, and Alexander Waibel, "An empirical exploration of CTC acoustic models," in *ICASSP*. IEEE, 2016, pp. 2623–2627.

[8] Haşim Sak, Andrew Senior, Kanishka Rao, Ozan Irsoy, Alex Graves, Françoise Beaufays, and Johan Schalkwyk, "Learning acoustic frame labeling for speech recognition with recurrent neural networks," in *ICASSP*. IEEE, 2015, pp. 4280–4284.

[9] Chuanying Niu, Jinsong Zhang, Xuesong Yang, and Yanlu Xie, "A study on landmark detection based on CTC and its application to pronunciation error detection," in *APSIPA ASC*. IEEE, 2017, pp. 636–640.

[10] Kenneth N Stevens, "Toward a model for lexical access based on acoustic landmarks and distinctive features," *Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1872–1891, 2002.

[11] Di He, Boon Pang Lim, Xuesong Yang, Mark Hasegawa-Johnson, and Deming Chen, "Improved ASR for under-resourced languages through multi-task learning with acoustic landmarks," in *Interspeech*. ISCA, 2018, pp. 2618–2622.

[12] Di He, Boon Pang P Lim, Xuesong Yang, Mark Hasegawa-Johnson, and Deming Chen, "Selecting frames for automatic speech recognition based on acoustic landmarks," *Journal of the Acoustical Society of America*, vol. 141, no. 5, pp. 3468–3468, 2017.

[13] Di He, Boon Pang Lim, Xuesong Yang, Mark Hasegawa-Johnson, and Deming Chen, "Acoustic landmarks contain more information about the phone string than other frames for automatic speech recognition with deep neural network acoustic model," *Journal of the Acoustical Society of America*, vol. 143, no. 6, pp. 3207–3219, 2018.

[14] John S Garofalo, Lori F Lamel, William M Fisher, Johnathan G Fiscus, David S Pallett, and Nancy L Dahlgren, "The DARPA TIMIT acoustic-phonetic continuous speech corpus CD-ROM," in *Linguistic Data Consortium*, 1993.

[15] Douglas B Paul and Janet M Baker, "The design for the Wall Street Journal-based CSR corpus," in *the Workshop on Speech and Natural Language*. Association for Computational Linguistics, 1992, pp. 357–362.

[16] Alex Graves and Navdeep Jaitly, "Towards end-to-end speech recognition with recurrent neural networks," in *International Conference on Machine Learning*. ACM, 2014, pp. 1764–1772.

[17] Sharlene A Liu, "Landmark detection for distinctive feature-based speech recognition," *Journal of the Acoustical Society of America*, vol. 100, no. 5, pp. 3417–3430, 1996.

[18] Mark Hasegawa-Johnson, James Baker, Sarah Borys, Ken Chen, Emily Coogan, Steven Greenberg, Amit Juneja, Katrin Kirchhoff, Karen Livescu, Srividya Mohan, et al., "Landmark-based speech recognition: Report of the 2004 Johns Hopkins summer workshop," in *ICASSP*. IEEE, 2005, pp. 213–216.

[19] Xiang Kong, Xuesong Yang, Mark Hasegawa-Johnson, Jeung-Yoon Choi, and Stefanie Shattuck-Hufnagel, "Landmark-based consonant voicing detection on multilingual corpora," *Journal of the Acoustical Society of America*, vol. 141, no. 5, pp. 3468–3468, 2017.

[20] Kenneth N Stevens, Sharon Y Manuel, Stefanie Shattuck-Hufnagel, and Sharlene Liu, "Implementation of a model for lexical access based on features," in *Second International Conference on Spoken Language Processing*. ISCA, 1992, pp. 499–502.

[21] Mark Hasegawa-Johnson, "Time-frequency distribution of partial phonetic information measured using mutual information," in *International Conference on Spoken Language Processing*. ISCA, 2000, pp. 133–136.

[22] Kenneth N Stevens, "Evidence for the role of acoustic boundaries in the perception of speech sounds," *Journal of the Acoustical Society of America*, vol. 69, no. S1, pp. S116–S116, 1981.

[23] John J McCarthy, "Feature geometry and dependency: A review," *Phonetica*, vol. 45, no. 2-4, pp. 84–108, 1988.

[24] Xavier Glorot and Yoshua Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *International Conference on Artificial Intelligence and Statistics*, 2010, pp. 249–256.

[25] Herve A Bourlard and Nelson Morgan, *Connectionist speech recognition: a hybrid approach*, vol. 247, Springer Science & Business Media, 2012.