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]University of Illinois at Urbana-Champaign, Urbana, IL, USA
[2]Novumind Inc, Santa Clara, CA, USA [3]Amazon Alexa Speech, Seattle, WA, USA

ABSTRACT

Connectionist temporal classification (CTC) training criterion provides an alternative acoustic model (AM) training strategy for automatic speech recognition in an end-to-end fashion. Although CTC criterion benefits acoustic modeling without needs of time-aligned phonetics transcription, it remains in need of efforts of tweaking to convergence, especially in the resource-constrained scenario. In this paper, we proposed to improve CTC training by incorporating acoustic landmarks. We tailored a new set of acoustic landmarks to help CTC training converge more quickly 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 faster and smoother significantly when they are augmented by acoustic landmarks. The models pretrained with mixed target labels can be finetuned furthermore, which reduced phone error rate by 8.72% on TIMIT. Consistent performance gain is also observed on reduced TIMIT and WSJ as well, in which case, we are the first to succeed testing the effectiveness of acoustic landmark theory on mid-sized ASR tasks.

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

1. INTRODUCTION

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

Training CTC models is not very stable and sometimes they are apt to converge to even a sub-optimal alignment very slowly, especially on resource-constrained data. In order to alleviate such common problem of CTC training, additional tricks are needed, for example, ordering training utterances by their lengths [6] or bootstrapping CTC models with CE trained models on fixed alignments [8]. The success of bootstrapping method with prior alignments indicates that external acoustic phonetics 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 locations of acoustic landmarks where abrupt manner changes of articulation exist [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] provides such fine-grained transcriptions (5.4 hours). The outcomes of testing these approaches are limited since the corpus is very small. It is worth to further explore the power of landmark theory when scaled up to large vocabulary continuous speech recognition (LVCSR) systems.

In this paper, we propose to augment phone sequences with acoustic landmarks for CTC acoustic modeling, and leverage a pretrain and finetune two-phase training procedure to address convergence problems of CTC training. 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 (PER) reduction in comparison to phone-based CTC baseline. We also conduct experiments on TIMIT with reduced size and WSJ. Our findings demonstrate that we are the

[*] Authors contributed equally.

first to succeed testing the generalization ability of acoustic landmark theory to the mid-sized ASR tasks.

2. BACKGROUND

2.1. Connectionist Temporal Classification (CTC)

Recent end-to-end systems have attracted much attention and they come with benefits of avoiding time-consuming iterations between alignments and model building. RNNs combined with the CTC loss function [3] are very popular to E2E systems [15]. The CTC loss computes the total likelihood of the target label sequence over all possible alignments given the input feature sequence, so that the computing is more expensive than frame-wise cross-entropy training. A blank symbol is introduced to augment the target label sequence to the same length with its input acoustic features. Forward-backward algorithms are used to make the likelihood summation over all possible alignments efficiently. The 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 $\mathbf{y}$ is the target label sequence, $\mathbf{x}$ is the input acoustic sequence, $\boldsymbol{\pi}$ is a blank-augmented sequence for $\mathbf{y}$, and $\mathcal{B}^{-1}(\mathbf{y})$ is the set of all such sequences. The by-product of relying on the indirect loss criterion is that the model requires more data and more training iterations to converge. During decoding, the target label sequences can be obtained by either greedy search or a WFST-based decoder. We report WERs based on WFST-based decoder.

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]. Landmark knowledge has been demonstrated to improve ASR performance [16, 17]. It is capable of reducing decoding complexity for DNN/HMM models [12, 13] and even achieves better recognition accuracy [11]. Previous works [11, 12, 13, 18] annotated landmark positions mostly following experimental findings presented in [19, 20]. 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 [21]. A collection of these binary features can distinguish each segment from all others in a languages. Auto-segmental phonology [22] 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 helps to distinguish major catogories of vowels, approximants, nasals, stops, fricatives, and affricates, such that another set of landmarks that focus on manner changes of articulation are defined. Table 1 illustrates the broad classes of sounds on TIMIT with respect to combination of two binary features. In the reduced phone set used by WSJ, phones from the top left category, from Table 1, is merges with respect phones in the bottom left set. In this case, only 3 types of categories are defined.

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 two broad classes of sounds even if manner changes don't exist. Figure 1 demonstrates an example of our two augmentation methods. A Landmark in *Mixed Label 2* usually represents a detailed boundary across two adjacent phones

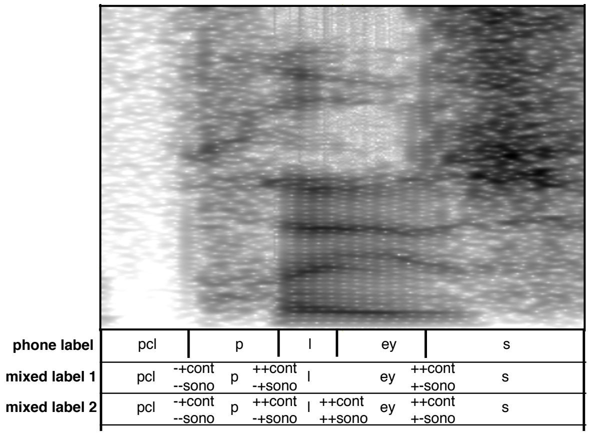

Fig. 1. Examples of target label sequences for the word "PLACE". The audio clip is selected from `SI792` on TIMIT.

where fine-grained manner transitions are pinpointed. CTC only requires a single target label sequence, so that augmenting phone sequences with landmarks can relax needs of time-aligned phone transcriptions. We believe *Mixed Label 2* works better since the manipulations of landmarks are related to the meaning of blank symbol in CTC training. The blank label in CTC was introduced to play two key roles in the training process [3]: first, it provides a filler or a default state for the model to fall back on when it is not confident enough to predict a non-blank label; second, mostly in character based models, it separates identical labels to indicate a repetition of the same label. We suspect, with experimental support, that the acoustic features of manner transition between phones are in fact easier to learn than phones themselves, contributing to a faster convergence rate.

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 sequence augmented by landmarks until convergence; at the phase of finetuning, the AM initializes weights from pretrained model and continues to be trained by label sequence with only phones. These two phases of training take the same training data. 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 only calculates over phones.

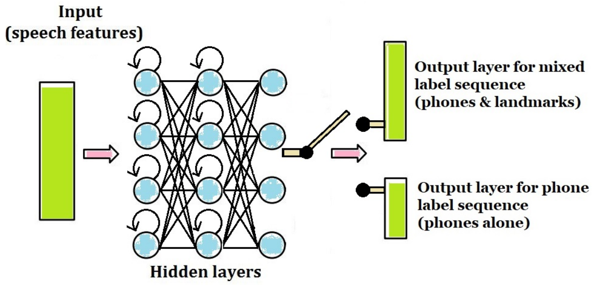
Input
(speech features)
Hidden layers
Output layer for mixed label sequence (phones & landmarks)
Output layer for phone label sequence (phones alone)

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 and WSJ [23] corpus. We used 40-dimensional log mel filterbank energy features computed with 10ms shift and 20ms span. No Delta feature or frame stacking was used. Hidden layers of our network begins with two bi-LSTM layers, each with 1024 cells (512 cells per direction), and we topped that with a fully connected layer with 256 neurons. Weights are initialized randomly following a uniform distribution. Our training flow enables new-bob annealing [24] after a minimum waiting period of 2 epochs. The initial learning rate is 0.0005. We trained our TIMIT baseline on 61 phones. The WSJ baseline is trained on 71 distinctive labels, originating from the 39 phones in the CMU set, similar to EESEN [5]. 1-best most likely phones are scored against the correct phone sequence to calculate the PER. No phone set reduction (mapping hypothesis sequence to the 39-count CMU set) was used before PER scoring. For both TIMIT and WSJ, we use the same train/dev/test split as Kaldi Recipe. We leverage the decoding graphs from EESEN (TGPR and TG) [1] to calculate WER for WSJ.

4.2. Experiments on TIMIT

Figure 3 presents the development set PER trend as the training epoch increases. The PER for mixed sequence represented by the red and yellow lines in Figure 3 is calculated when all landmarks are removed from the label sequence. When the model is trained with the mixed label sequence, no error rate advantage, against the baseline, can be observed. However, the model converges much more rapidly and smoothly. After the model converges on the mixed label sequence, we retrain it on another output layer with phone labels only. The retrain, represented by a purple line in Figure 3, returns a model that is more accurate, and the system also reaches convergence more rapidly than the baseline.

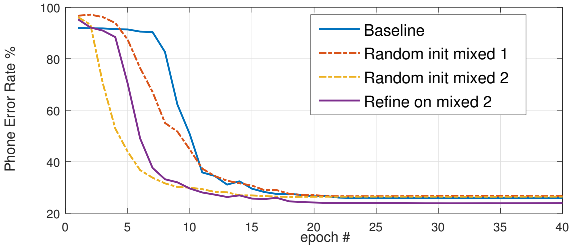

Fig. 3. PER with respect to epoch number (all PER reported on phones only: landmarks are stripped away if necessary)

After the network is trained on mixed labels, we refine the networks on an output layer with phone labels only. 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 is reasonably high among past works, such as [3], calculating PER directly on the full TIMIT phone set. As we can see, if we train with mixed labels and strip away landmarks from the hypothesis sequence, landmarks do not provide any benefit in terms of absolute PER. In fact, maybe due to the model being distracted by landmarks, the PER tends to be slightly higher. However, the model can achieve lower PER in the finetuning

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

stage when landmark labels are removed. A relative phone error rate reduction of 4.64% and 8.72% can be observed for the two labeling methods mentioned in Section 3.1.

It is not clear to us why we need a refine stage to achieve more outstanding accuracy. But we suspect there might be imperfection in the landmark labels that resulted in none ideal error rates. We looked into the count distribution, 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, ordered in category permutation using continuant as the first variable and sonorant as the second), based on method 2 mentioned in Section 3.1, in TIMIT. It seems distribution of landmarks cannot be considered balance. Most labels indicate a transition related to the +*continuant*+*sonorant* phones. Imbalance in the distribution of landmark count is not ideal for augmenting phone recognizer training as it tends to provide the same and redundant information for a large portion of the cases.

Table 2. Comparison between baseline and our proposed models with augmented target labels in PER (%). Number in the parenthesis 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%)

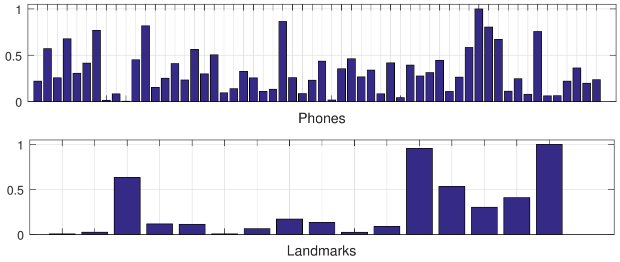

Fig. 4. Prior distributions of phones and acoustic landmarks.

4.3. Beyond Standard TIMIT

To further evaluate the effectiveness of our method of refining mixed labels. We stretched the training data into two different directions. In this section, we only consider method 2 mentioned in Section 3.1 as it clearly out-performs the other empirically. Also, only error rate for the refine model is reported. On one end, we reduced training data from the standard training set of TIMIT. In 5, we randomly held back some utterance from joining the training set.

As illustrated in Figure 5, when we reduce the percentage of training data used beyond a point, both the refined model and baseline fails to converge. Observing the output sequence revels the model starts to output a constant sequence (usually made up of two to three frequent phones) for all utterance. As the network fails to move out of this state, the new-bob annulling continues to reduce the learning rate and the illy formed network becomes frozen. However, Figure 5 shows that the refined flow can still train a reasonable model, with PER no larger than 30% on the development set, when the baseline fails to converge. This can be beneficial in cases when training data is very limited.

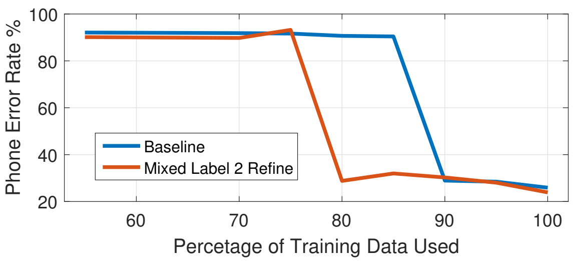

Fig. 5. Performance by varying the amount of training data.

On the other end, we experimented our flow on WSJ. Compared to EESEN [5], our baseline is slightly underperforming. However, our network is more shallower, we also did not include Deltas in our feature. A performance gap can be anticipated. As we can see from Table 3, although the accuracy margin is not as outstanding as TIMIT, the refined model still out-performs the baseline consistently across all metric. To our best knowledge, this is the first work where the effective of landmark is evaluated on a mid-sized corpus.

Table 3. Error Rate (Percentage) on WSJ

	PER		WER (tgpr / tg)	
Test set	eval92	dev93	eval92	dev93
random init	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 acoustic modeling (AM) with Acoustic Landmark. We modified the classic landmark definition to suit the CTC criterion and implemented a two-phase training flow to improve CTC AMs. Experiments on TIMIT and WSJ demonstrate that CTC training becomes more stable and rapid when phone label sequences are augmented by landmarks. Our methods also help CTC training achieve lower error rate by a large margin, up to 8.72%. More importantly, the advantage is consistent across both corpus and all metrics. We also found the flow can train a converging model with very limited training data while the baseline fails. In addition, without the need of time-alignments, we demoed the effectiveness of Acoustic Landmark theory on WSJ, a mid-size corpus with no phone duration information.

6. REFERENCES

[1] Wayne Xiong, Jasha Droppo, Xuedong Huang, Frank Seide, Mike Seltzer, Andreas Stolcke, Dong Yu, and Geoffrey Zweig, "Achieving human parity in conversational speech recognition," *arXiv preprint arXiv:1610.05256*, 2016.

[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," *arXiv preprint arXiv:1703.02136*, 2017.

[3] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *ICML 2006*. 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 2015*, 2015.

[5] Yajie Miao, Mohammad Gowayyed, and Florian Metze, "EESEN: End-to-end speech recognition using deep RNN models and WFST-based decoding," in *ASRU 2015*. 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 *ICML 2016*, 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 2016*. 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 2015*. 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 2017*. IEEE, 2017, pp. 636–640.

[10] Kenneth N Stevens, "Toward a model for lexical access based on acoustic landmarks and distinctive features," *J. 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 2018*. ISCA, 2018.

[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," *J. 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," *J. 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 cdrom," in *Linguistic Data Consortium*, 1993.

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

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

[17] 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 2005*. IEEE, 2005, vol. 1, pp. I–213.

[18] Xiang Kong, Xuesong Yang, Mark Hasegawa-Johnson, Jeung-Yoon Choi, and Stefanie Shattuck-Hufnagel, "Landmark-based consonant voicing detection on multilingual corpora," *arXiv preprint arXiv:1611.03533*, 2016.

[19] 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*, 1992.

[20] Mark Hasegawa-Johnson, "Time-frequency distribution of partial phonetic information measured using mutual information," in *6th International Conference on Spoken Language Processing*, 2000.

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

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

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

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