IMPROVED ASR FOR UNDER-RESOURCED LANGUAGES THROUGH MULTI-TASK LEARNING WITH ACOUSTIC LANDMARKS

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

[1]Coordinated Science Lab, University of Illinois at Urbana-Champaign, Urbana, Illinois, USA 61801
[2]Novumind Inc, Santa Clara, USA 95054
[3]Beckman Institute, University of Illinois at Urbana-Champaign, Urbana, Illinois, USA 61801

ABSTRACT

Acoustic Landmarks have been shown to improve frame-synchronous HMM-based speech recognition when applied as a secondary channel or in a post-processing step. Landmarks are, furthermore, somewhat language-universal, therefore their characteristics can potentially be learned once from a suitably labeled corpus and rapidly applied to many other languages in a scalable fashion. Earlier works tend to use approaches that add-on significantly to the overall complexity of the system. This work proposes using Landmarks as the secondary task in multi-task learning, thereby obtaining the benefits of Landmark without increasing decoder complexity. Our experiments demonstrate improvement, up to 3%, in both a well-resourced source language (English) and an under-resourced adaptation target language (Iban), thereby supporting the hypothesis that Landmarks have useful and complementary information to phone labels, both monolingually and cross-lingually.

Index Terms— Acoustic Landmarks, Under-resourced ASR, Multi-task Learning

1. INTRODUCTION

Acoustic Landmark theory [1, 2] suggests that there exist abrupt changes and local extrema in both speech articulation and spectrogram, that these key acoustic time-points offer unusually high information density about the phonetic content of the signal, and that speech perception uses the salience of these events to speed up the process of speech understanding. These instantaneous acoustic events, and the surrounding acoustic cues that label them, are called acoustic landmarks. Studies like [3, 4] have shown that classic frame-synchronous ASR can benefit from Landmark information, but often [3] at the cost of expensive additional computation. This work considers improving ASR accuracy through the help of acoustic Landmarks, by improving acoustic model (AM) quality at the training phrase through Multi-task Learning (MTL) [5].

MTL has shown the ability to improve the performance of speech models, especially those based on neural networks [6, 7, 8]. In MTL, if the secondary task complements the primary task, there is a chance that the jointly trained model offers higher accuracy [6]. Landmark detection could potentially be an ideal secondary task for automatic speech recognition (ASR; Fig 1), since it detects instantaneous events that are informative to phone recognition. Also, because they are based on universal properties of speech production and speech perception, Landmarks are likely to be more useful for cross-language ASR adaptation [9] and are more noise-immune [3, 10] than other secondary tasks that have been used in MTL. These characteristics are especially helpful for under-resourced languages: in an under-resourced language, training data may be limited, e.g., there may be little or even no transcribed speech. A Landmark-based system trained on a well-resourced language might be adapted to an under-resourced language, thus improving ASR accuracy in the under-resourced language.

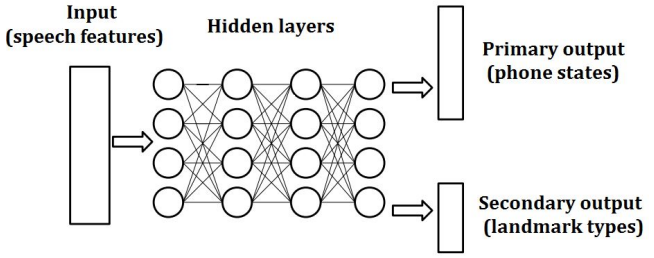

Fig. 1. MTL Neural Network Jointly Trained on Phone States and Landmark Types

The key contributions of this work are experimental findings supporting the hypothesis that acoustic Landmark information can reduce ASR error rate through MTL. This finding is consistent across two corpora (TIMIT [11] and Iban [12]) and both mono- and tri-phone AM. We also observed evidence that acoustic Landmark detectors can be used to aid MTL cross-lingually, in the sense that our Landmark detector trained in English seems to benefit an Iban ASR. Key methodology and techniques used to apply the Landmark theory to MTL are explained in Sec. 3. Results are presented in Sec. 4, and the paper concludes and discusses future work in Sec. 5.

2. BACKGROUND

The theory of acoustic Landmarks assumes that the phonemes of every language are derived from distinctive features: approximately-binary speech sound categories defined by properties of speech perception [13], or production [1], or statistically frequent patterns of phonology [14]. Distinctive features are perceived through acoustic cues scattered in speech. The locations at which these cues occur in speech are called acoustic Landmarks. Speech processing systems using distinctive features have multiple advantages over phoneme based systems. As opposed to phonemes, which are language dependent, distinctive features are more universal. This advantage offers the latter stronger portability across languages [9]. In addition, distinctive-feature-based systems tend to be more noise-immune [3, 10]. It has been shown [3, 4] that phone-based ASR can benefit from the use of landmark and distinctive feature information, e.g., in a side channel, or in a post-processing phase. More details on Landmark definition and their location can be found in Sec 3, where we label the Landmarks according to human annotated phone boundaries from the TIMIT corpus.

2.1. Multi-task Learning

Multi-task Learning (MTL) [5] has shown the ability to improve statistical model performance by jointly training a single model for multiple purposes. The multiple tasks in MTL share the same input, but generate multiple outputs predicting likelihoods for a primary and one or more secondary tasks. When the multiple tasks are related but not identical, or (in the ideal case) complementary to each other, MTL models offer higher accuracy [6]. A number of works [6, 7, 8] have proved MTL to be effective on speech processing tasks. Among them [8] proved MTL effective at improving model performance for under-resourced ASR.

Most works mentioned above [6, 7, 8] leverage the flexibility of Deep Neural Network (DNN) based classifiers with Softmax output layers. The output of a typical Softmax DNN, $P_c(x)$, with C distinctive classes, is calculated by Eq 1, where x is the input feature, and y_i represents the ith output from the immediate preceding layer, right after the affine transform, without applying a non-linear operation.

$$P_c(x) = \frac{exp(y_c)}{\sum_{i=1}^{C}(exp(y_i))}, \forall c = 1...C \tag{1}$$

Experiments in this paper employed cross-entropy as the cost for network training. The sum of cross-entropies, $\mathcal{L}$, between each individual output and their expected value, l_i, as in Eq 2, is back-propagated to update the neuron weights.

$$\mathcal{L} = \sum_{i=1}^{C}(l_i \log(P_i(x))) \tag{2}$$

When we conduct MTL, for the same input x, we prepare two sets of labels. The label l_i^{ph} specifies the phone or triphone state associated with a frame, while l_j^{la} encodes the presence and type of acoustic Landmark. Eq 2 becomes Eq 3, where α is a trade-off value we use to weight the two sets of labels. We sweep through a small list of candidate α's to find the value that returns the best result on development test data.

$$\mathcal{L}_{mtl} = (1-\alpha)\sum_{i=1}^{C^{ph}}(l_i^{ph} \log(P_i^{ph}(x))) + \alpha \sum_{j=1}^{C^{la}}(l_j^{la} \log(P_j^{la}(x))) \tag{3}$$

2.2. The Iban Corpus

Training under-resourced speech processing systems can be challenging [8]. Landmark based speech processing systems, compared to classic frame-synchronous systems, are potentially more promising to these tasks, because landmarks are more portable across languages than phonemes [9] and more noise-immune than static phonetic segments [3, 10]. For these reasons we form the hypothesis that Landmark-based models trained on resource-rich languages can be more effectively transfered to under-resourced languages than pure phone-based models.

The under-resourced language studied in this paper is Iban [12]. Iban is a language spoken in Borneo, Sarawak (Malaysia), Kalimantan and Brunei. Malay tends to have more clearly articulated consonants than most other Asian and European languages (according to our own informal perceptual survey of many languages). We chose Iban for experiments because, if Iban consonants are articulated as clearly as those of Malay, then we hypothesize that an Iban ASR might gain particular benefit from a Landmark detector, even if the Landmark detector is trained in a completely unrelated language (English). The Iban corpus contains 8 hours of clean speech from 23 speakers, 17 speakers contributed $6.8h$ of training data, and the test-set contain $1.18h$ of data from 6 speakers. The language model was trained on a $2M$-word Iban news dataset using SRILM [15].

3. METHODOLOGY

Landmark definitions in this paper, listed in Table 1, are based primarily on those of [16], with small modifications. Modifications include the elimination of the +33% and -20% offsets after the beginning or before the end of some phones, reported in [16] and [17], in favor of the simpler definitions in Table 1.

We extracted Landmark training labels by referencing the TIMIT human annotated phone boundaries. An example of the labeling is presented in Fig 2. This example from [4] illustrates the labeling of the word "Symposium"[1]. The figure is generated using Praat [18].

[1] selected from audio file: `TIMIT/TRAIN/DR1/FSMA0/SX361.WAV`

Table 1. Landmark types and their positions for acoustic segments, where 'c', and 'r' denote consonant closure, and release; 'start', 'middle', and 'end' denote three positions across acoustic segments, respectively.

Manner of Articulation	Landmark Type and Position
Vowel	V: middle
Glide	G: middle
Fricative	Fc: start, Fr: end
Affricate	Sr,Fc: start, Fr: end
Nasal	Nc: start, Nr: end
Stop Closure	Sc: start, Sr: end

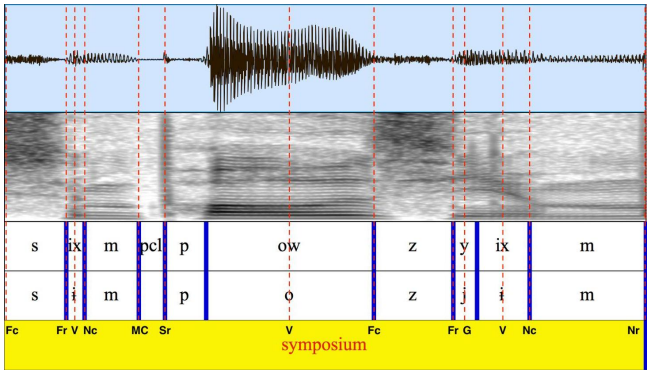

Fig. 2. Acoustic landmark labels for the pronunciation of word "Symposium".

3.1. Bootstrapping TIMIT for a Landmark Detector

All ASR systems in this paper use the feature extraction methods proposed in [7]. No speaker adaptation is used in any of the ASR systems in this paper, because it is not yet clear how to perform speaker adaptation for Landmark detectors; we reserve that problem for future papers. The network is initialized using Deep Belief Network (DBN) [19] pre-training.

When applying the Landmark labels to MTL, we did encounter difficulties. We failed to realize that our main goal was to train a Landmark detector that can effectively compliment the phone state recognizer, not to train a Landmark detector that can optimally detect Landmark locations. An MTL that over-emphasizes the Landmark detection criterion tends to perform poorly as an ASR AM, because Landmarks are relatively infrequent compared to phone-state-labeled speech frames: every frame has a phone label, but fewer than 20% of frames have a Landmark label. Because of the sparsity of Landmark-labeled frames, weighting the MTL criterion to emphasize Landmark accuracy increased the number of frames receiving the same label, "No Landmark," and reduced the benefit of Landmark detection as a secondary task for MTL. We explored different ways to adjust the Landmark labels. Table 2 covers some of these adjustments. When we label the Landmark on only the frame in which it occurs ($\mathbf{ver}_1$), the MTL AM returns higher WER than the baseline in Iban. Expanding the range of a Landmark to include the nearby 2 frames ($\mathbf{ver}_2$) returned the best result. The third labeling ($\mathbf{ver}_3$) expanded the Landmark region, but split the center frame and nearby frames into different classes. $\mathbf{ver}_4$ marked Landmark labels similarly to $\mathbf{ver}_3$, but distinctly labeled frames before vs. after the Landmark. Expanding the domain of the Landmark was helpful ($\mathbf{ver}_2$), but separate classes for frames far from the Landmark ($\mathbf{ver_3}$ and $\mathbf{ver}_4$) seemed to be less helpful. To further address the imbalance among different Landmark classes, the training criterion was computed using a weighted sum of training data, with weights inversely proportion to the class support.

Table 2. *Iban Tri-phone WER Comparison of Different Landmark Labeling Techniques*

Baseline	$\mathbf{ver}_1$	$\mathbf{ver}_2$	$\mathbf{ver}_3$	$\mathbf{ver}_4$
18.24	18.31	18.01	18.16	18.11

3.2. Cascading the MTL to Iban

After we trained a Landmark detector on TIMIT, we ran the detector on Iban. The English-trained Landmark detector output is used to define reference labels for the secondary task of the Iban acoustic model MTL. An example of the detector output on an arbitrary utterance[2] in Iban is given in Fig 3. We found that the results are good at outlining Fricative landmarks. The detector can also find Stop Closure Landmarks near the correct locations, but with less precision than the Fricative Landmarks. The performance on Vowel and Glide Landmarks is only fair: the detector often mixes up the two classes, and incorrectly labels non-Vowels as Vowels.

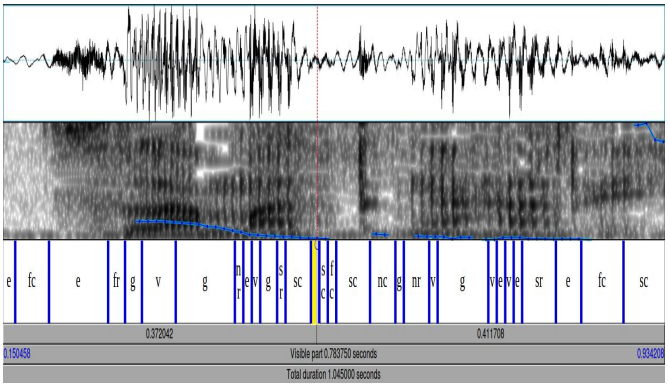

Fig. 3. Landmark Detection Result on Iban for utterance ibm_003_049, pronouncing **selamat tengah ari** (**s-aa-l-a-m-a-t t-aa-ng-a-h a-r-i** in Iban phone set). Transcription labels: e=empty (no Landmark); fr, fc, sr, sc, nr, nc, v, g are as in Table 1.

We experimented with multiple ways to initialize the Landmark detector and the phone recognizer in the second

[2] `iban/data/wav/ibm/003/ibm_003_049.wav`

language. We found that using a network trained through MTL in TIMIT to initialize the MTL network in the second language yields the best results. This is similar to to cascading phone state recognition and Landmark detection MTL in Iban after the same tasks are done in (TIMIT) English. We found the technique marginally but consistently outperforms other initializations including DBN.

4. RESULTS

All experiments were conducting using the Kaldi [20] toolbox. We extracted 40-dimensional log-mel-filter-bank (FBank) features, and concatenated them with their delta and double-delta coefficients, then spliced 11 consecutive frames from the surrounding context to form an input vector for each frame (10ms shift, 25ms span). The AM is a deep neural network with 4 hidden, fully-connected layers, 2048 nodes/layer. The same features and network structure were used for both the Landmark detector, the MTL model and the baseline. The baseline is initialized using a DBN.

Results are reported in Table 3 for both English (TIMIT) and Iban. The main goal of this study is to examine whether Landmark-based MTL is useful cross-lingually; TIMIT results are reported to indicate the performance of Landmark-based MTL in the source language, prior to cross-language adaptation. Similar tradeoffs were observed in both TIMIT and Iban with respect to model tuning and parameter selection. On development test sets in both corpora, the value $\alpha = 0.2$ returned the lowest error rate (with little variability in the range $0.1 \leq \alpha \leq 0.3$), and was therefore used for evaluation. Higher values of α (higher weight for the Landmark labels) resulted in higher ASR error rates. The Landmark detector achieves 80.11% frame-wise accuracy in validation, which is lower than Landmark detection accuracies reported in some other studies, apparently because the Landmark criterion was de-emphasized by MTL. Conversely, phone error rate (PER) was reasonably good: 20.6% for the baseline system, and 20.0% for the MTL system, as compared to 22.7% for the open-source Kaldi tri4_nnet recipe.

Decoding results for Iban are reported using Word Error Rate (WER), because the Iban corpus is distributed with automatic but not manual phonetic transcriptions. The comparison between PER in TIMIT and WER in Iban permits us to demonstrate that Landmark-based MTL can benefit PER in a source language (English), and WER in an adaptation target language (Iban). The triphone-based ASR trained without MTL on TIMIT, then adapted to Iban, achieves 18.24% WER; a system that is identical but for the addition of Landmark-based MTL can achieve 18.01% WER. Neither system includes speaker adaptation, and therefore neither system is better than the 17.45% state of the art WER for this corpus[3].

[3] https://github.com/kaldi-asr/kaldi/blob/master/egs/iban/s5/RESULTS

Table 3. *Decoding Error Rate for mono-phone (Mono) and tri-phone (Tri) on TIMIT and Iban.*

Corpus	TIMIT (PER)		Iban (WER)	
Acoustic Model	Mono	Tri	Mono	Tri
Baseline	24.6	20.6	24.58	18.24
MTL	24.2	20.0	24.11	18.01

As we can see in Table 3, in all cases, regardless of AM and corpus, the ASR system jointly trained with Landmark and phone information returns lower error rate. The PER reduction on TIMIT is greater than the WER reduction on Iban, yet since the Landmark MTL models consistently return lower WER across different AMs, we find the tendency promising.

5. DISCUSSION AND FUTURE WORK

This study confirms that acoustic Landmarks convey information complementary to phone recognition, as has been shown in previous work [3, 4], and confirms that Landmark detectors can be trained in one language and applied in another, as has been shown by [9]. This study is the first to demonstrate the use of Landmark detection as the secondary task in MTL, and to demonstrate a consistent resulting drop in ASR error rates, in both a well-resourced source language and an under-resourced target language.

While a cross-language Landmark detector provides useful information complementary to the orthographic transcription, visual inspection indicates that a cross-language Landmark detector is not as accurate as a same-language Landmark detector. Future work, therefore, will train a more accurate Landmark detector, using recurrent neural network methods that do not depend on human-annotated phone boundaries, and that can therefore be more readily applied to multilingual training corpora.

6. ACKNOWLEDGEMENTS

This research was partially supported by the Qatar National Research Fund (QNRF) grant 7-766-1-140.

7. REFERENCES

[1] Kenneth N. Stevens, "Evidence for the role of acoustic boundaries in the perception of speech sounds," in *Phonetic Linguistics: Essays in Honor of Peter Ladefoged*, Victoria A. Fromkin, Ed., pp. 243–255. Academic Press, Cambridge MA USA, Orlando, Florida, 1985.

[2] Kenneth N Stevens, "Acoustic phonetics," Cambridge, 2000, pp. 1–615, MIT Press, Cambridge MA USA.

[3] 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, vol. 1, p. 1213.

[4] 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," *arXiv preprint arXiv:1710.09985*, 2017.

[5] Rich Caruana, *Multitask Learning*, pp. 95–133, Springer US, Boston, MA, 1998.

[6] Z. Wu, C. Valentini-Botinhao, O. Watts, and S. King, "Deep neural networks employing multi-task learning and stacked bottleneck features for speech synthesis," in *ICASSP*, April 2015, pp. 4460–4464.

[7] M. L. Seltzer and J. Droppo, "Multi-task learning in deep neural networks for improved phoneme recognition," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, May 2013, pp. 6965–6969.

[8] D. Chen, B. Mak, C. C. Leung, and S. Sivadas, "Joint acoustic modeling of triphones and trigraphemes by multi-task learning deep neural networks for low-resource speech recognition," in *ICASSP*, 2014, pp. 5592–5596.

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

[10] Katrin Kirchhoff, "Combining articulatory and acoustic informatino for speech recognition in noisy and reverberant environments," in *ICSLP*, 1998, pp. 0873:1–4.

[11] 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.

[12] Sarah Samson, Laurent Besacier, Benjamin Lecouteux, and Mohamed Dyab, "Using Resources from a Closely-related Language to Develop ASR for a Very Under-resourced Language: A Case Study for Iban," in *Interspeech 2015*, Dresden, Germany, Sept. 2015.

[13] Roman Jakobson, C Gunnar Fant, and Morris Halle, "Preliminaries to speech analysis. the distinctive features and their correlates.," 1951, pp. 1–64, MIT Press, Cambridge MA USA.

[14] Noam Chomsky and Morris Halle, "The sound pattern of english," 1968, pp. 1–484, MIT Press, Cambridge MA USA.

[15] Andreas Stolcke, "Srilm–an extensible language modeling toolkit," in *ICSLP*, 2002.

[16] Sharlene A Liu, "Landmark detection for distinctive feature-based speech recognition," in *The Journal of the Acoustical Society of America*, 1996, vol. 100, pp. 3417–3430.

[17] Mark Hasegawa-Johnson, "Time-frequency distribution of partial phonetic information measured using mutual information," in *ICSLP*, 2000, pp. 133–136.

[18] Paul Boersma and D Weenik, "Praat: a system for doing phonetics by computer. report of the institute of phonetic sciences of the university of amsterdam," *Amsterdam: University of Amsterdam*, 1996.

[19] Abdel-rahman Mohamed, George E Dahl, and Geoffrey Hinton, "Acoustic modeling using deep belief networks," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 14–22, 2012.

[20] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al., "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society, 2011, number EPFL-CONF-192584.