Interspeech 2018
2-6 September 2018, Hyderabad

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

dihe2@illinois.edu, bplim@novumind.com, xyang45@illinois.edu, jhasegaw@illinois.edu, dchen@illinois.edu

Abstract

Furui first demonstrated that the identity of both consonant and vowel can be perceived from the C-V transition; later, Stevens proposed that acoustic landmarks are the primary cues for speech perception, and that steady-state regions are secondary or supplemental. Acoustic landmarks are perceptually salient, even in a language one doesn't speak, and it has been demonstrated that non-speakers of the language can identify features such as the primary articulator of the landmark. These factors suggest a strategy for developing language-independent automatic speech recognition: landmarks can potentially be learned once from a suitably labeled corpus and rapidly applied to many other languages. This paper proposes enhancing the cross-lingual portability of a neural network by using landmarks as the secondary task in multi-task learning (MTL). The network is trained in a well-resourced source language with both phone and landmark labels (English), then adapted to an under-resourced target language with only word labels (Iban). Landmark-tasked MTL reduces source-language phone error rate by 2.9% relative, and reduces target-language word error rate by 1.9%-5.9% depending on the amount of target-language training data. These results suggest that landmark-tasked MTL causes the DNN to learn hidden-node features that are useful for cross-lingual adaptation.

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

1. Introduction

In the early 1980s, Furui [1] demonstrated that the identity of both consonant and vowel can be perceived from a 100ms segment of audio extracted from the C-V transition; in 1985, Stevens [2] proposed that acoustic landmarks are the primary cues for speech perception, and that steady-state regions are secondary or supplemental. Acoustic landmarks produce enhanced response patterns on the mammalian auditory nerve [3], and it has been demonstrated that non-speakers of a language can identify features such as the primary articulator of the landmark [4]. Automatic speech recognition (ASR) systems have been proposed that depend completely on landmarks, with no regard for the steady-state regions of the speech signal [5], and such systems have been demonstrated to be competitive with phone-based ASR under certain circumstances. Other studies have proposed training two separate sets of classifiers, one trained to recognize landmarks, another trained to recognize steady-state phone segments, and fusing the two for improved accuracy [6] or for reduced computational complexity [7, 8]. It has been difficult to build cross-lingual ASR from such systems, however, because very few of the world's languages possess large corpora with the correct timing of consonant release and consonant closure landmarks manually coded. In this paper we propose a different strategy: we propose to use reference landmark labels in only one language (the source language). A landmark detector trained in the source language is ported to the target language in two ways: (1) by automatically detecting landmark locations in target language test data, and (2) by using landmark detection as a secondary task for the purpose of training a triphone state recognizer that can be more effectively ported cross-lingually. The neural network is trained with triphone state recognition as its primary task; landmarks are introduced as a secondary task, using the framework of multi-task learning (MTL) [9].

MTL has shown the ability to improve the performance of speech models, especially those based on neural networks [10, 11, 12, 13]. MTL is a mechanism for reducing generalization error. A single-task neural net is provably optimal, for large enough training datasets: as the size of the training dataset goes to infinity, if the number of hidden nodes is set equal to the square root of the number of training samples, the difference between the network error rate and the Bayes error rate goes to zero [14]. MTL is useful when the training dataset is too small to permit zero-error learning [10], or when the training dataset and the test dataset are drawn from slightly different probability distributions (e.g., different languages). In either case, MTL proposes training the network to perform two tasks simultaneously. The secondary task is not important during test time, but if the network is forced to perform the secondary task during training, it will sometimes learn network weights (and consequently, hidden layer activation functions) that are either (1) less prone to over-fitting on the training data than a single-task network, or (2) generalize better from the distribution of the training data to the distribution of the test data. 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. Because landmarks have been demonstrated to correlate with non-linguistic perceptual signals (e.g., enhanced response on the auditory nerve [3]) and because features of a landmark can be classified by non-speakers of the language [4], it is possible that the secondary task of landmark detection and classification will force a neural net to learn weights that are more useful for cross-language ASR adaptation [15] than those of a single-task network. 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

10.21437/Interspeech.2018-1124

on a well-resourced language might be adapted to an under-resourced language, thus improving ASR accuracy in the under-resourced language. Furthermore, we carried out experiments reducing the training data in the secondary language, examining the effectiveness of Landmark detection as a secondary task for MTL in very low resourced (40 minutes) scenarios. To our best knowledge, this is the first study where Acoustic Landmark has been applied to under-resource ASR training.

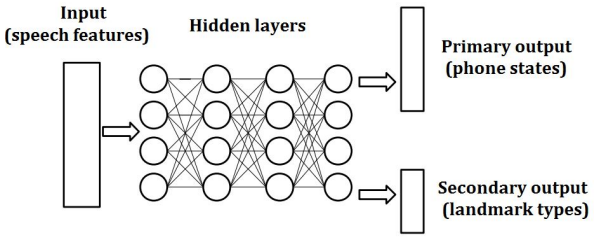

Figure 1: *MTL Neural Network Jointly Trained on Phone States and Landmark Types*

The work is presented as follows, after we review some background in Sec. 2, 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 in Sec. 5.

2. Background

Before we talk about our methodology, we would like to briefly review MTL as a neural network training method and talk about the under-resource corpus we used in this study.

2.1. Multi-task Learning

Multi-task Learning (MTL) [9] 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 better generalization from training to test corpus [10]. A number of works [10, 11, 12] have proved MTL to be effective on speech processing tasks. Among them [12] proved MTL effective at improving model performance for under-resourced ASR.

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. The network is trained in order to minimize, on the training data, a multi-task error metric as shown in Eq. 1, where $P_i^{ph}(x)$ $(1 \leq i \leq C^{ph})$ is the probability of monophone or triphone state i at frame x as estimated by the neural network, $P_j^{la}(x)$ $(1 \leq j \leq C^{la})$ is the probability of landmark label j at frame x as estimated by the network, and 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.

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

2.2. The Iban Corpus

The under-resourced language studied in this paper is Iban [16]. Iban is a language spoken in Borneo, Sarawak (Malaysia), Kalimantan and Brunei. The Malay phone set is similar to English, e.g., the two languages have the same inventory of stop consonants and affricates; Malay also has a relatively transparent orthography, in the sense that the pronunciation of a word is usually well predicted by its written form. If Iban orthography is as transparent as Malay, and if its phone set is as similar to English (an approximated mapping between the Iban phone set and IPA can be found at [1]), then it is possible that a landmark detector trained on English may perform well in Iban. However, we are not trying to claim Malay or Iban is a perfect secondary language, when compared to English, for our experiments. These languages are different in many aspects, for example, English in particular is notable for its consonant clustering and use of diphthongs and even triphthongs; this is not the case in Malay. Iban is also selected because of the recent release of an Iban training and test corpus with particularly good quality control [16]. The Iban corpus contains 8 hours of clean speech from 23 speakers. Seventeen speakers contributed $6.8h$ of training data, and the test-set contains $1.18h$ of data from 6 speakers. The language model was trained on a $2M$-word Iban news dataset using SRILM [17]. We foresee that if the primary and under-resource languages share more similarities than English and Iban, we have a good chance of observing better results than what we have obtained.

3. Methods

We trained an ASR on the TIMIT corpus using the methods of multi-task learning (Sec. 2.1), using the detection and classification of landmarks (Sec. 3.1) as a secondary task. The same ASR is then adapted cross-lingually to the Iban corpus (Sec. 2.2)

3.1. Defining and Marking Landmarks

Landmark definitions in this paper, listed in Table 1, are based primarily on those of [18], 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 [18] and [19], in favor of the simpler definitions in Table 1.

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

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 [8] illustrates the labeling of the word "Symposium"[2]. The figure is generated

[1] https://github.com/dihe2/interspeech18/blob/master/phone_mapping.txt

[2] selected from file: TIMIT/TRAIN/DR1/FSMA0/SX361.WAV

using Praat [20].

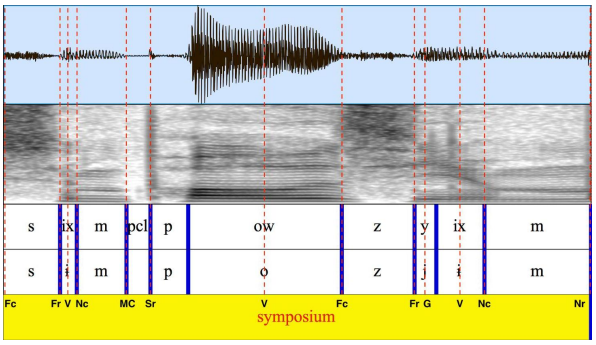

Figure 2: *Acoustic landmark labels for the pronunciation of word "Symposium".*

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, we explored different ways to adjust the Landmark labels to achieve the best MTL performance. We found, expanding the range of a Landmark to include the nearby 2 frames returns the highest accuracy for the primary task.

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 [21].

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 [3] 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 sonorant consonants as vowels.

When applying the landmark detector to Iban, we are concerned with the error generated by the detector. The automatically detected landmark labels are treated as ground truth for MTL in landmark-task MTL in Iban, therefore it is possible that erroneously detected landmarks may mis-lead the network training. To minimize the effect of these mistakes, we introduce an extra weighting factor in the MTL training criterion based on the confidence of the landmark detector output, as shown in Eq. 2.

$$\mathcal{L}_x = (1 - \alpha c_x) \sum_{i=1}^{C^{ph}} (l_i^{ph} \log(P_i^{ph}(x))) + \alpha c_x \sum_{j=1}^{C^{la}} (l_j^{la} \log(P_j^{la}(x))) \quad (2)$$

where c_x is a confidence value derived based on the landmark detector output for feature frame x based on Eq 3.

$$c_x = P_m^{la\_de}(x) - \frac{1}{C^{la} - 1} \sum_{k=1, k \neq m}^{C^{la}} (P_k^{la\_de}(x)) \quad (3)$$

where $P_i^{la\_de}(x)$ is the softmax output for landmark class i. The class index $m = \text{argmax}_i^{C^{la}} P_i^{la\_de}(x)$, which is also the index for the class the landmark detector predicted.

The intuition behind this extra layer of weighting is to assign a penalty, during training of the ASR, that is proportional to our certainty of its error. If the detector is not confident separating the output class from other classes, then we reduce the loss it generates in the MTL process.

We experimented with multiple ways to initialize the landmark detector and the phone recognizer in the second 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. We found the technique marginally but consistently outperforms other initializations including Deep Belief Networks (DBN) [22].

4. Results

All experiments were conducting using the Kaldi [23] toolbox. We extracted an acoustic feature vector using the same algorithm and parameters as [11]. The acoustic model (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 [22]. No speaker adaptation is used in any of the ASR systems in this paper.

Results are reported in Table 2 for both English (TIMIT) and Iban. TIMIT results are reported to indicate the performance of Landmark-based MTL in the source language, prior to cross-language adaptation.

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. For larger α values, such as $\alpha > 0.4$ the WER starts to drop significantly. Error rate higher than the baseline starts to appear, for some setups, when $\alpha \geq 0.6$. The landmark detector achieves 80.11% frame-wise accuracy in validation. 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). Triphone-based ASR trained without MTL on TIMIT, then adapted to Iban, achieves 18.4% WER; a system that is identical but for the addition of landmark-task MTL can achieve 17.93% WER. Neither system includes speaker adaptation, and therefore neither system is better than the 17.45% state of the art WER for this corpus[4] with the same language model.

As we can see in Table 2, in all cases, regardless of AM and corpus, the ASR system jointly trained with landmark and

[3] iban/data/wav/ibm/003/ibm_003_049.wav

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

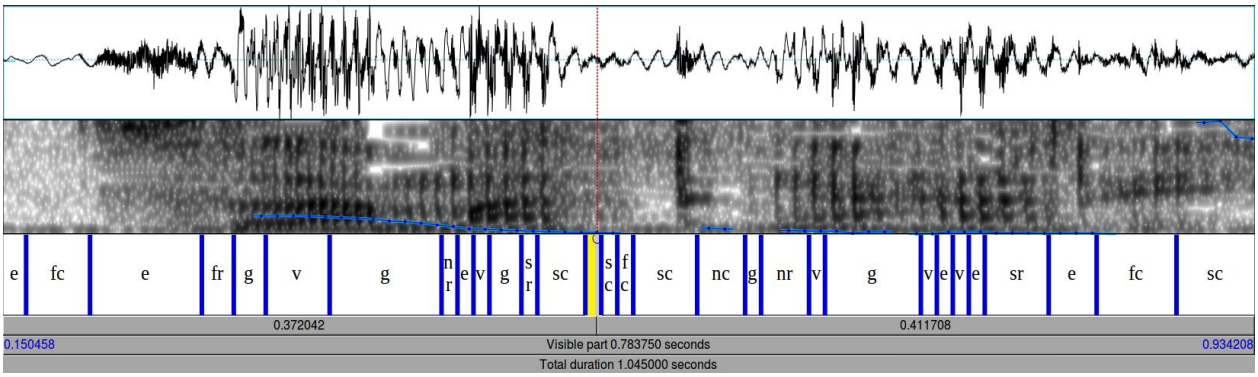

Figure 3: *Landmark Detection Result on Iban for utterance ibm_003_049, pronouncing* **selamat tengah ari** *(***s-aa-l-a-m-a-t t-aa-ng-a-h a-r-i** *in Iban phone set). Transcription labels: e=empty (no Landmark); fr, fc, sr, sc, nr, nc, v, g are as in Table 1.*

Table 2: *Decoding Error Rate for mono-phone (Mono) and tri-phone (Tri) on TIMIT and Iban.*

Corpus	AM	Baseline	MTL	MTL w/ Confid
TIMIT (PER)	Mono	24.6	24.2	NA
	Tri	20.6	20.0	NA
Iban-full (WER)	Mono	24.62	24.22	24.18
	Tri	18.40	18.03	17.93
Iban-25% (WER)	Mono	28.87	27.97	27.64
	Tri	21.31	20.70	20.63
Iban-10% (WER)	Mono	31.16	28.49	28.48
	Tri	25.12	23.64	23.57

phone information returns lower error rate. The setups "Iban-25%" and "Iban-10%" train the AM on only 25% (100 minutes) and 10% (40 minutes) of the training data uniformly selected at random from the Iban training set (maintaining speaker and gender ratio), but evaluates the error rate on the full test set. As the amount of training data decreases, the benefits of MTL increase. When only 10% of training data is available, simulating a very low resource case, MTL reduces the word error rate by the greatest margin: 8.7% for monophone ASR and 6.17% for triphone ASR. Weighting the MTL loss according to confidence results in a small but consistent error rate reduction. All systems use the same language model, and all systems use acoustic models with the same network architecture and feature set; the error rate change we observe is caused entirely by the use of landmark-task MTL. We foresee that the difference between English and Iban may have some negative effect on the experimental results, and that 2 languages that share more similarities may benefit from our approach even more.

5. Discussion and Future Work

This demonstrates that landmark-task MTL results in a neural network that can be more effectively ported cross-lingually. As the amount of training data in the under-resourced language is reduced (from 400 minutes to 100 or 40 minutes), the benefits of landmark-task MTL increase. In addition, introducing a loss weighting according the landmark detector confidence seems to reduce the effect of landmark detector error as it consistently produces lower error rate.

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 multi-lingual training corpora.

6. Acknowledgements

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

7. References

[1] Sadaoki Furui, "On the role of spectral transition for speech perception," *J. Acoustical Society of America*, vol. 80, no. 4, pp. 1016–1025, 1983.

[2] 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.

[3] Bertrand Delgutte and Nelson Y.S. Kiang, "Speech coding in the auditory nerve: IV. Sounds with consonant-like dynamic characteristics," *J. Acoustical Society of America*, vol. 75, pp. 897–907, 1984.

[4] Preethi Jyothi and Mark Hasegawa-Johnson, "Acquiring speech transcriptions using mismatched crowdsourcing," in *Proc. AAAI*, 2015, pp. 1263–1269.

[5] Amit Juneja and Carol Espy-Wilson, "A novel probabilistic framework for event-based speech recognition," *J. Acoustical Society of America*, vol. 114, no. 4(A), pp. 2395, 2003.

[6] 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.

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

[8] 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," *The Journal of the Acoustical Society of America*, vol. 143, no. 6, pp. 3207–3219, 2018.

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

[10] 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.

[11] 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.

[12] 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.

[13] Xuesong Yang, Kartik Audhkhasi, Andrew Rosenberg, Samuel Thomas, Bhuvana Ramabhadran, and Mark Hasegawa-Johnson, "Joint modeling of accents and acoustics for multi-accent speech recognition," *arXiv preprint arXiv:1802.02656*, 2018.

[14] Andrew R. Barron, "Approximation and estimation bounds for artificial neural networks," *J. Machine Learning*, vol. 14, pp. 115–133, 1994.

[15] 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.

[16] 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.

[17] Andreas Stolcke, "SRILM—an extensible language modeling toolkit," in *ICSLP*, 2002.

[18] 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.

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

[20] 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.

[21] Xuesong Yang, Anastassia Loukina, and Keelan Evanini, "Machine learning approaches to improving pronunciation error detection on an imbalanced corpus," in *Spoken Language Technology Workshop (SLT), 2014 IEEE*. IEEE, 2014, pp. 300–305.

[22] 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.

[23] 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.