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

Multi-Task Learning using Mismatched Transcription for Under-Resourced Speech Recognition

Van Hai Do[1,4], *Nancy F. Chen*[2], *Boon Pang Lim*[2], *Mark Hasegawa-Johnson*[1,3]

[1]Viettel Group, Vietnam
[2]Institute for Infocomm Research, A*STAR, Singapore
[3]University of Illinois at Urbana-Champaign, USA
[4]Advanced Digital Sciences Center, Singapore

haidovan@gmail.com, {nfychen, bplim}@i2r.a-star.edu.sg, jhasegaw@illinois.edu

Abstract

It is challenging to obtain large amounts of native (matched) labels for audio in under-resourced languages. This could be due to a lack of literate speakers of the language or a lack of universally acknowledged orthography. One solution is to increase the amount of labeled data by using mismatched transcription, which employs transcribers who do not speak the language (in place of native speakers), to transcribe what they hear as nonsense speech in their own language (e.g., Mandarin). This paper presents a multi-task learning framework where the DNN acoustic model is simultaneously trained using both a limited amount of native (matched) transcription and a larger set of mismatched transcription. We find that by using a multi-task learning framework, we achieve improvements over monolingual baselines and previously proposed mismatched transcription adaptation techniques. In addition, we show that using alignments provided by a GMM adapted by mismatched transcription further improves acoustic modeling performance. Our experiments on Georgian data from the IARPA Babel program show the effectiveness of the proposed method.

Index Terms: mismatched transcription, probabilistic transcription, multi-task learning, low resourced languages

1. Introduction

[1] There are more than 6700 languages spoken in the world today *(www.ethnologue.com)*, but only a few of them have been studied by the speech recognition community. Almost all academic publications describing ASR in a language outside the "top 10" are focused on the same core research problem: the lack of transcribed speech training data to build the acoustic model. Various methods have been proposed for acoustic modeling of under-resourced languages. They are summarized in four main groups.

The first group is based on a universal phone set [1, 2] that is generated by merging phone sets of different languages according to the international phonetic alphabet (IPA). After that a multilingual acoustic model can be trained for all languages using the common phone set.

The idea of the second group such as cross-lingual subspace Gaussian mixture models (SGMMs) [3, 4] and multilingual deep neural networks (DNNs) [5–7] is to create an acoustic model that can be effectively broken down into two components in which the main component captures language-independent statistics and the other component captures language specific statistics.

In the third group, the well-resourced language (i.e., source language) acoustic model is used as a feature extractor to generate high-level features such as source language phone posteriors for the target language speech data. These features enable the use of simpler models for the target language. Several examples of this approach are cross-lingual tandem [8, 9], cross-lingual Kullback-Leibler based HMM (KL-HMM) [10, 11], phone mapping [12–14], and exemplar-based modeling [15,16].

The fourth group is mismatched crowdsourcing which was recently proposed as a potential approach to deal with the lack of native transcribers to produce labeled training data [17–22]. In this approach, the transcribers do not speak the under-resourced language of interest (target language), they write down what they hear in this language into nonsense syllables in their native language (source language) called mismatched transcription. This mismatched transcription is then converted by a mismatched channel model into target language transcription in a lattice format called probabilistic transcription (PT). PT is then used to adapt existing acoustic models which can be GMM [19] or DNN [20]. One disadvantage of this approach is that we rely on the quality of the mismatched channel to convert mismatched transcription to probabilistic transcription of the target language. And hence information can be lost in this process. Moreover, the mismatched channel can only be trained using limited parallel training data given the lack of linguistic resources and/or native speakers i.e., same audio segments with both matched and mismatched transcriptions which results in a poor converting performance.

In this paper, we propose a method to use mismatched transcription directly in a multi-task learning framework without the need of parallel training data. Specifically, a DNN acoustic model is trained using two softmax layers, one for matched transcription and one for mismatched transcription. The motivation using a DNN with multiple softmax layers is from previous studies for multi-task learning (MTL) [23–25] and multilingual DNN training [5–7]. In [23], MTL was used for noise robust speech recognition for a digit recognition task. Given the observed noisy speech, the neural network was trained to predict both the digit label and the clean speech. In [24], MTL was applied for continuous phoneme recognition where a multiple softmax layer DNN was trained to predict not only context independent phonetic states but phoneme identity and phonetic context. Experiments on TIMIT showed a consistent improvement was achieved by using MTL. In audio visual speech

[1]This study is supported by the research grant for the Human-Centered Cyber-physical Systems Programme at the Advanced Digital Sciences Center from Singapore's Agency for Science, Technology and Research (A*STAR)

Copyright © 2017 ISCA

http://dx.doi.org/10.21437/Interspeech.2017-788

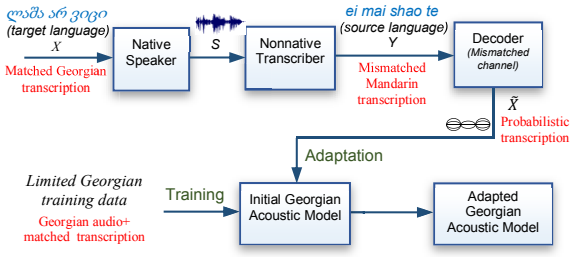

Figure 1: *Mismatched channel model for mismatched crowdsourcing: the target language is Georgian, while the source language is Mandarin.*

recognition, applying MTL is a natural approach. In [25], a 2-softmax layer DNN was used for an audio visual ASR where one softmax layer is used for audio data and the second one is for visual data. Shared-hidden-layer multilingual DNN (SHL-DNN) introduced in [5] can be viewed as a variant of MTL using a DNN with multiple softmax layers for different languages trained simultaneously and benefit from each other. Our MTL-DNN architecture is similar to the SHL-DNN in [5]. However in our MTL-DNN, the input is the audio feature of the target language while the 2 softmax output layers are used to predict target language (matched) transcription and source language (mismatched) transcription. After training, the MTL-DNN can model speech perception of native listeners and foreign listeners. This framework provides a natural structure to share speech perception of source and target language listeners on the target language speech.

Our proposed technique can be viewed as a hybrid of multilingual DNN [5] and audio-visual ASR techniques [25]. The audio-visual example demonstrates that DNNs can learn an embedding that captures the related structure of two different types of parallel data while the multilingual examples show that DNNs can learn a shared representation for similar data with disjoint ground truth labels. The proposed technique is a combination of the two: the objective of the DNNs is to learn the structure embedded in the incorrect transcriptions (e.g., lexical mappings or word segmentations) that improve performance in the disjoint but similar target data.

The rest of this paper is organized as follows: Section 2 gives a brief introduction of mismatched transcription and its application in speech recognition. Section 3 presents our proposed MTL-DNN framework. Experiments are shown in Section 4. Conclusion is presented in Section 5.

2. Mismatched Channel Model for Mismatched Crowdsourcing

Mismatched crowdsourcing was proposed to solve the shortage of native transcription in under-resourced languages [22]. As shown in Figure 1, the input to the system is a message, X, in the under-resourced utterance language, which is realized as a speech signal S. Transcribers from the source language listen to S, and write nonsense syllables, Y, in the orthography of the annotation language; Y is called the mismatched transcription. A decoder trained with limited parallel data is used to estimate X given Y [17, 18]. Estimated $\tilde{X}$ is normally in the lattice format called probabilistic transcription (PT). PT is then used to adapted the initial acoustic model trained with limited target language data [19, 20, 26].

3. Proposed Multi-task Learning Architecture

There are two limitations of the process in Figure 1. First, the mismatch channel is probabilistic: the maximum a posteriori transcription is not guaranteed to be correct. Second, the mismatch channel model is only trained using limited parallel training data, i.e., audio with both matched and mismatched transcription; since there is little such audio, the channel model is under-trained. This paper proposes a method to use mismatched transcription directly in a multi-task learning (MTL) framework. As shown in Figure 2, a MTL-DNN acoustic model has two softmax layers, one for matched (target language - Georgian) transcription and one for mismatched (source language - Mandarin) transcription. Georgian frame alignment is given by forced alignment using the initial Georgian GMM trained with limited Georgian data as in the conventional DNN training procedure. To obtain frame alignment for the mismatched transcription, we introduce a GMM mismatched acoustic model trained using the target language (Georgian) audio data with source language (Mandarin) mismatched transcription (Figure 3). After training, the mismatched GMM acoustic model is used to do forced alignment on the adaptation set to achieve frame alignment for DNN training. With the proposed approach, we do not need to use parallel corpus to train the mismatched channel.

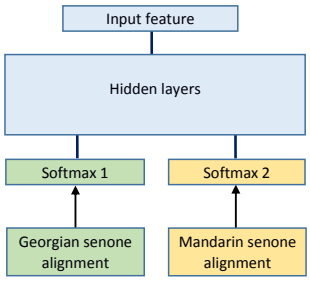

Figure 2: *Multi-task learning DNN framework using both matched and mismatched transcription.*

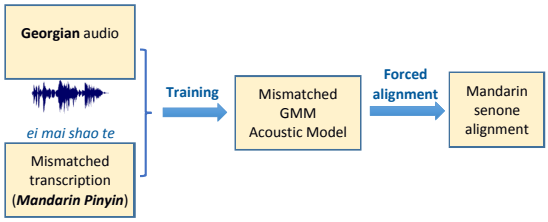

Figure 3: *Target language (Georgian) audio and mismatched transcription (Mandarin Pinyin) are used to build the mismatched acoustic model.*

In this paper, cross-entropy is used as the objective function to train the MTL-DNN. Since we have two softmax layers, two corresponding cross-entropy functions are used, they are

Eq (1) for softmax 1 (Georgian)

$$J_1 = -\sum_t \sum_i \hat{y}_{1i}(t) \log y_{1i}(t) \quad (1)$$

where $y_{1i}(t) \in [0, 1]$ is the value of the i^{th} output of the softmax layer 1 at time t, $\hat{y}_{1i}(t) \in \{0, 1\}$ is the training label at time t given by forced alignment of the matched GMM acoustic model, and i is a Georgian triphone state (senone).

Table 1: *Phone Error Rate (PER %) for different acoustic models with different amounts of matched training transcription.*

Training data	w/o PT adaptation		w/ PT adaptation	
	GMM	DNN	GMM	DNN
12 minutes	73.71	73.66	72.16	71.67
24 minutes	71.52	70.90	69.94	69.74
48 minutes	69.88	69.54	68.75	68.70

Eq (2) for softmax 2 (Mandarin)

$$J_2 = -\sum_t \sum_k \hat{y}_{2k}(t) \log y_{2k}(t) \tag{2}$$

where $y_{2k}(t) \in [0, 1]$ is the value of the k^{th} output of the softmax layer 2 at time t, $\hat{y}_{2k}(t) \in \{0, 1\}$ is the training label at time t given by forced alignment of the mismatched GMM acoustic model, and k is a Mandarin trigrapheme state (senone).

In this paper, the MTL-DNN is trained to minimize the following multi-task objective function.

$$J = p_2 J_1 + \alpha p_1 J_2 \tag{3}$$

where p_1, p_2 are the priors of training data size for matched and mismatched training data, respectively. p_1, p_2 are added to deal with data imbalance between two datasets. α is the combination weight for the mismatched softmax layer. When $\alpha = 0$, the MTL-DNN becomes a conventional DNN using only one Georgian softmax layer.

After the MTL-DNN is trained using both matched and mismatched transcriptions, the softmax layer for mismatched transcription is discarded. We only keep the softmax layer for matched transcription (target language) for decoding as in the conventional single-task DNN.

4. Experiments

The IARPA BABEL Georgian corpus (IARPA-babel404bv1.0a data release) provided in the context of the 2016 NIST Open Keyword Search Evaluation is used for our experiments. The acoustic data are collected from various real noisy scenes and telephony conditions. For the pronunciation lexicon, 1-letter graphemes are used to approximate phonemes.

In our experiments, Georgian is chosen as the under-resourced language and Mandarin speakers are chosen as non-native transcribers. We randomly select 12, 24 and 48 minutes from the 3-hr very limited language pack set (VLLP) with native transcription to simulate limited transcribed training data conditions. Together, 10 hours from the untranscribed portion of the training data were chosen by maximizing the number of speakers via choosing 70 seconds from the middle of each Georgian conversation. A total of 4 Mandarin transcribers were hired from Upwork (https://www.upwork.com/), each in charge of 2.5 hrs. Each transcriber listened to short Georgian speech segments and wrote down transcription in Pinyin alphabet that is acoustically closest to what he thinks he heard [22].

To convert mismatched transcription to matched transcription, a mismatched channel is modeled as a finite memory process using weighted finite state transducer (WFST). Mismatched channel represents the non-native transcriber, who hears Georgian phones, and generates Mandarin Pinyin graphemes; the channel is decoded using a MAP decoder to generate Georgian phones from Mandarin orthography. The weights on the arcs of the WFST model are learned using the EM algorithm [27] to maximize the likelihood of the observed training instances. The USC/ISI Carmel finite-state toolkit [28] is used for EM training of the WFST model and the OpenFST toolkit [29] is used for all finite-state operations. The Kaldi speech recognition toolkit [30] is used to build speech recognition. Input feature is Mel filter bank energies plus F0 (pitch), acoustic models are GMM with speaker adaptive training (SAT) and DNN. During the decoding process, a bigram phonetic language model trained from training transcription is used. Performance of all the systems are evaluated in phone error rate (PER) on 20 minutes extracted from the 10-hour development set given by NIST.

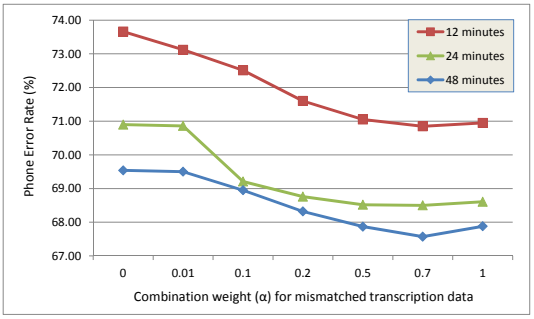

Figure 4: *Phone error rate versus combination weight α of mismatched transcription in the multi-task learning framework for the case of 10 hours mismatched transcription.*

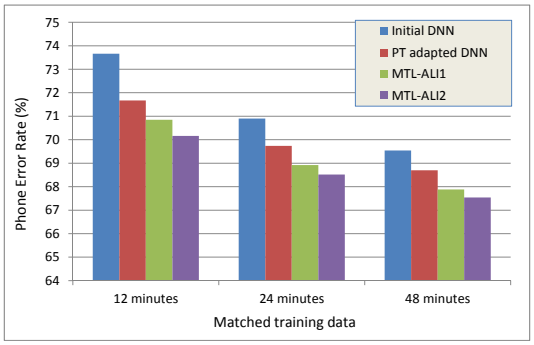

Figure 5: *PER given by different DNN acoustic models for the case of 12, 24 and 48 minutes of matched training data.*

4.1. Baselines

We first investigate performance of Georgian phone recognition when only limited amounts of transcribed training data are available to train the acoustic models. The first two columns of Table 1 show PER for the GMM and DNN for the case of 12, 24 and 48 minutes of Georgian training data sizes. We can see that when more training data are available, PER reduces consistently and using DNN acoustic model results in a small improvement over GMM. The main reasons for these high PERs are: the corpus is noisy telephone conversational speech and the training data are very limited. To investigate the usefulness of mismatched transcription, we first decode the mismatched channel to convert mismatched transcription to probabilistic transcription (PT) and then use PT to adapt the existing acoustic models (Figure 1). Two adaptation approaches are conducted:

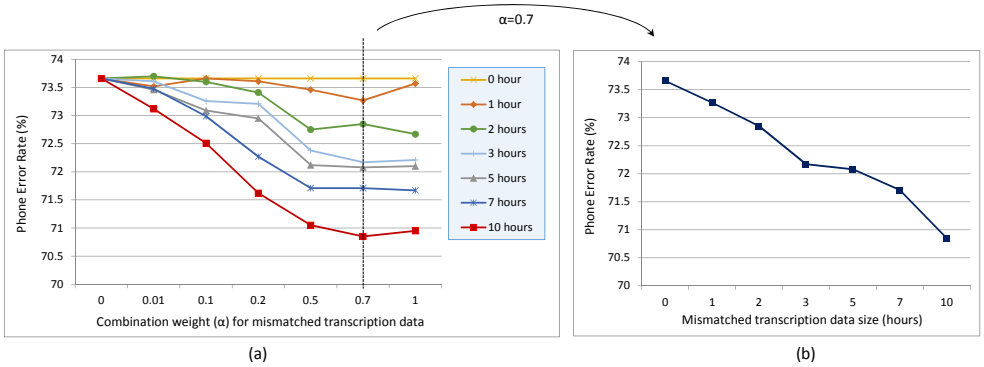

Figure 6: *PER given by MTL-ALI1 with different amounts of mismatched data for the case of 12 minutes matched training data. (a) combination weight α runs from 0 to 1; (b) combination weight $\alpha = 0.7$.*

- GMM is adapted using MAP adaptation [19].
- DNN is adapted by further training with probabilistic transcription [20].

As shown in the last two columns of Table 1, using PT adaptation improves both the GMM and DNN acoustic models. However, this gain saturates when we have more Georgian data to train the initial acoustic models i.e., 48 minutes.

4.2. Multi-task learning

Figure 4 shows PER given by the proposed MTL framework (Figure 2) for the case of 12, 24 and 48 minutes of matched transcription. The combination weight α for the mismatched transcription data is varied from 0 to 1. When $\alpha = 0$, this is the case of conventional monolingual DNN with only one matched data softmax layer as shown in the second column of Table 1. When α increases, we can see that the MTL framework can consistently improve performance for all three cases. There is not much difference when α runs from 0.5 to 1. When $\alpha = 0.7$, we achieve the best performance with 70.85%, 68.50%, 67.57% PER for the case of 12, 24, 48 minutes of matched transcription, respectively. These results are better than PER given by the GMM and DNN models with and without PT adaptation in Table 1.

In the above experiments, frame alignment for the matched output layer of the DNN is provided by the initial GMM trained with limited training data (i.e., the first column of Table 1). However, Table 1 shows that by using PT adaptation for the GMM, we obtain consistent improvement. Our hypothesis is that this better GMM can result in better alignment for DNN training which will benefit the performance of our MTL framework. Figure 5 illustrates PER given by different DNN acoustic models. The first bars in each of the three groups are the initial DNNs trained only by limited amounts of matched transcription data. The second bars are the results given by the DNNs after PT adaptation. The third and fourth bars are the PER given by our MTL-DNN using alignment given by the initial GMM and GMM with PT adaptation, respectively (with $\alpha = 0.7$). It can be seen that both the two MTL-DNNs outperform the initial DNN and DNN with PT adaptation. In addition, using GMM with PT adaptation to generate frame alignment results in a consistent improvement over using the initial GMM. It shows that our hypothesis is correct. This can be considered as two-level-adaptation where the first adaptation level is for GMM to generate better alignment and the second level is for DNN to improve acoustic modelling. Figure 5 also shows that when less matched training data are available for the target language, we achieve larger improvement using mismatched transcription. This means our approach is especially suitable for extremely under-resourced languages.

4.3. Effect of adaptation data size on MTL

In this section, we investigate how mismatched transcription data size affects MTL performance. Figure 6 illustrates the PER given by MTL using different mismatched transcription data sizes while matched Georgian data size is 12 minutes. In this case, the alignment for the Georgian output layer is provided by the initial monolingual GMM (thus the point at 10 hours with $\alpha = 0.7$, 70.8% PER, matches the third bar of Figure 5, the case of 12 minutes). PER is shown to drop consistently when more mismatched transcription data are available for MTL.

4.4. Discussion

While Georgian and Mandarin are not considered to be languages that share many similarities, we still observe consistent gains when we use Mandarin mismatched transcriptions. We are currently investigating language pairs that are closer to each other, such as Mandarin (source language) and Singapore Hokkien (under-resourced target language).

In this paper, a MTL-DNN with two softmax layers was used. Obviously, MTL-DNN can consist of more than two softmax layers, which allows MTL-DNN to be able to simultaneously use different types of mismatched transcription such as from multiple source languages.

5. Conclusion

We proposed a multi-task learning framework to improve speech recognition for under-resourced languages. Specifically, the MTL-DNN acoustic model is simultaneously trained using both a limited amount of native (matched) transcription and a larger set of mismatched transcription. Experiments conducted on the IARPA BABEL Georgian corpus showed that by using the proposed method, we achieve consistent improvements over monolingual baselines and previously proposed mismatched transcription adaptation techniques. Moreover, we proposed a two-level-adaptation technique where alignments provided by the GMM adapted by mismatched transcription is used to further improve MTL-DNN performance. In this paper, we also investigated that using more mismatched transcription data results in a consistent improvement.

6. References

[1] T. Schultz and A. Waibel, "Experiments on Cross-Language Acoustic Modeling," in *ICSLP*, 2001, pp. 2721–2724.

[2] N. T. Vu, F. Kraus, and T. Schultz, "Cross-language bootstrapping based on completely unsupervised training using multilingual A-stabil," in *ICASSP*, 2011.

[3] L. Burget, P. Schwarz, M. Agarwal, P. Akyazi, K. Feng, A. Ghoshal, O. Glembek, N. Goel, M. Karafiát, D. Povey *et al.*, "Multilingual acoustic modeling for speech recognition based on subspace Gaussian mixture models," in *ICASSP*, 2010, pp. 4334–4337.

[4] L. Lu, A. Ghoshal, and S. Renals, "Maximum a posteriori adaptation of subspace Gaussian mixture models for cross-lingual speech recognition," in *ICASSP*, 2012, pp. 4877–4880.

[5] J. T. Huang, J. Li, D. Yu, L. Deng, and Y. Gong, "Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers," in *ICASSP*, 2013, pp. 7304–7308.

[6] N. T. Vu, D. Imseng, D. Povey, P. Motlicek, T. Schultz, and H. Bourlard, "Multilingual deep neural network based acoustic modeling for rapid language adaptation," in *ICASSP*, 2014, pp. 7639–7643.

[7] H. Xu, V. H. Do, X. Xiao, and E. S. Chng, "A comparative study of BNF and DNN multilingual training on cross-lingual low-resource speech recognition," in *INTERSPEECH*, 2015, pp. 2132–2136.

[8] A. Stolcke, F. Grezl, M.-Y. Hwang, X. Lei, N. Morgan, and D. Vergyri, "Cross-domain and cross-language portability of acoustic features estimated by multilayer perceptrons," in *ICASSP*, 2006, pp. 321–324.

[9] P. Lal and S. King, "Cross-lingual automatic speech recognition using tandem features," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 12, pp. 2506–2515, 2013.

[10] D. Imseng, H. Bourlard, and P. N. Garner, "Using KL-divergence and multilingual information to improve ASR for under-resourced languages," in *ICASSP*, 2012, pp. 4869–4872.

[11] D. Imseng, P. Motlicek, H. Bourlard, and P. N. Garner, "Using out-of-language data to improve an under-resourced speech recognizer," *Speech communication*, vol. 56, pp. 142–151, 2014.

[12] K. C. Sim and H. Li, "Context-sensitive probabilistic phone mapping model for cross-lingual speech recognition." in *INTERSPEECH*, 2008, pp. 2715–2718.

[13] V. H. Do, X. Xiao, E. S. Chng, and H. Li, "Context-dependent phone mapping for LVCSR of under-resourced languages," in *INTERSPEECH*, 2013, pp. 500–504.

[14] V. H. Do, X. Xiao, E. S. Chng, , and H. Li, "Cross-lingual phone mapping for large vocabulary speech recognition of under-resourced languages," in *IEICE Transactions on Information and Systems*, vol. E97-D, no. 2, 2014, pp. 285–295.

[15] T. N. Sainath, B. Ramabhadran, D. Nahamoo, D. Kanevsky, D. V. Compernolle, K. Demuynck, J. F. Gemmeke, J. R. Bellegarda, , and S. Sundaram, "Exemplar-based processing for speech recognition: An overview," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 98–113, 2012.

[16] V. H. Do, X. Xiao, E. S. Chng, and H. Li, "Kernel Density-based Acoustic Model with Cross-lingual Bottleneck Features for Resource Limited LVCSR," in *INTERSPEECH*, 2014, pp. 6–10.

[17] P. Jyothi and M. Hasegawa-Johnson, "Acquiring speech transcriptions using mismatched crowdsourcing," in *AAAI*, 2015, pp. 1263–1269.

[18] P. Jyothi and M.Hasegawa-Johnson, "Transcribing continuous speech using mismatched crowdsourcing," in *INTERSPEECH*, 2015, pp. 2774–2778.

[19] C. Liu, P. Jyothi, H. Tang, V. Manohar, R. Sloan, T. Kekona, M. Hasegawa-Johnson, and S. Khudanpur, "Adapting ASR for under-resourced languages using mismatched transcriptions," in *ICASSP*, 2016, pp. 5840–5844.

[20] A. Das and M. Hasegawa-Johnson, "An investigation on training deep neural networks using probabilistic transcriptions," in *INTERSPEECH*, 2016, pp. 3858–3862.

[21] V. H. Do, N. F. Chen, B. P. Lim, and M. Hasegawa-Johnson, "Analysis of mismatched transcriptions generated by humans and machines for under-resourced languages," in *INTERSPEECH*, 2016, pp. 3863–3867.

[22] M. A. Hasegawa-Johnson, P. Jyothi, D. McCloy, M. Mirbagheri, G. M. di Liberto, A. Das, B. Ekin, C. Liu, V. Manohar, H. Tang *et al.*, "ASR for Under-Resourced Languages From Probabilistic Transcription," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 1, pp. 50–63, 2017.

[23] S. Parveen and P. D. Green, "Multitask learning in connectionist ASR using recurrent neural networks," in *EUROSPEECH*, 2003.

[24] M. Seltzer and J. Droppo, "Multi-task learning in deep neural networks for improved phoneme recognition," in *ICASSP*, 2013, pp. 6965–6969.

[25] A. Thanda and S. M. Venkatesan, "Multi-task Learning Of Deep Neural Networks For Audio Visual Automatic Speech Recognition," *arXiv preprint arXiv:1701.02477*, 2017.

[26] V. H. Do, N. F. Chen, B. P. Lim, and M. Hasegawa-Johnson, "Speech recognition of under-resourced languages using mismatched transcriptions," in *IALP*, 2016, pp. 112–115.

[27] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the royal statistical society. Series B (methodological)*, pp. 1–38, 1977.

[28] "Carmel finite-state toolkit." [Online]. Available: http://www.isi.edu/licensed-sw/carmel/

[29] C. Allauzen, M. Riley, J. Schalkwyk, W. Skut, and M. Mohri, "A general and efficient weighted finite-state transducer library," 2007.

[30] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The Kaldi speech recognition toolkit," in *ASRU*, 2011.