Acoustic Modeling for Under-resourced Language using Mismatched Transcriptions

Van Hai Do[1,2], Nancy F. Chen[3], Boon Pang Lim[4],
Mark Hasegawa-Johnson[5]
[1]Thuyloi University, Vietnam, [2]Viettel Group, Vietnam,
[3]Institute for Infocomm Research, Singapore, [4]Novumind, USA
[5]University of Illinois Urbana-Champaign, USA
haidv@tlu.edu.vn, nfychen@i2r.a-star.edu.sg, bplim@novumind.com,
jhasegaw@illinois.edu

Abstract

Mismatched crowdsourcing is a technique to derive speech transcriptions using crowd-workers unfamiliar with the language being spoken. This technique is especially useful for under-resourced languages since it is hard to hire native transcribers. In this paper, we demonstrate that using mismatched transcription for adaptation improves performance of speech recognition under limited matched training data conditions. We show that using previously published methods for training data augmentation improves the utility of mismatched transcription. Finally, we show that a mismatched transcription can be used to train one neural network in two forms, in two sequential steps: first as a probabilistic transcription, and second as the auxiliary task of a multi-task learner.

Keywords

speech recognition, mismatched transcription, under-resourced language, data augmentation, multi-task learning.

1. Introduction

Commercial automatic speech recognition (ASR) is available in fewer than one percent of the world's living languages (e.g., www.google.com/intl/en/chrome/demos/speech.html; arxiv.org/abs/1412.5567). Almost all academic publications describing ASR in a language outside the "one percent" are focused on the same core research problem: the lack of transcribed speech training data. Usually, to build a reasonable speech recognition system,

tens to hundreds of hours of training data are required, while commercial systems normally use thousands of hours of training data. This large resource requirement limits the development of a full fledged acoustic model for under-resourced languages. To deal with this issue, various methods have been proposed. They are summarized in four categories.

The first category is based on a universal phone set (Schultz et al., 2011; Vu et al., 2011) that is generated by merging phone sets of different languages according to the international phonetic alphabet (IPA) scheme. A multilingual acoustic model can therefore be trained for all languages using the common phone set.

In the second category, the idea is to create an acoustic model that can be effectively broken down into two parts in which the major part captures language-independent statistics and the other part captures language specific statistics. Typical examples in this approach are the cross-lingual subspace Gaussian mixture models (SGMMs) (Burget et al., 2011) and multilingual DNN (Xu et al., 2015).

In the third category, the source acoustic model acts as a feature extractor to generate cross-lingual features such as source language phone posteriors for the target language speech data. As these features are higher-level features as compared to conventional features such as MFCCs, they enable the use of simpler models trained with a small amount of training data to model the target acoustic space. Several examples of this approach are cross-lingual tandem (Stolcke et al., 2006), cross-lingual Kullback-Leibler based HMM (KL-HMM) (Imseng et al., 2012), phone mapping (Sim et al., 2008; Do et al., 2013; Do et al., 2014a), and exemplar-based models (Sainath et al., 2012; Do et al., 2014b).

The fourth category is mismatched crowdsourcing which was recently proposed as a potential approach to deal with the lack of native transcribers to produce labeled training data (Jyothi and Hasegawa-Johnson., 2015a; Jyothi and Hasegawa-Johnson., 2015b; Liu et al., 2016; Das and Hasegawa-Johnson., 2016; Do et al., 2016; Hasegawa-Johnson et al., 2017). In this method, the transcribers do not speak the under-resourced language of interest (target language), they write down what they hear in this language as nonsense syllables in their native language (source language). These transcriptions are "mismatched" because the source and target languages differ. The mismatched transcriptions are then converted by a channel decoder into target language transcription in a lattice format called probabilistic transcription (PT). PT is then used to adapt existing acoustic models which can either be GMM (Liu et al., 2016) or DNN (Das and Hasegawa-Johnson., 2016).

In this paper, we follow the fourth approach where mismatched transcription is used to improve performance of under-resourced speech recognition. Specifically, Vietnamese is chosen as the under-resourced language, and the mismatched transcription is generated by Mandarin speakers. In this study, we assume there is a limited amount of matched

transcription in the target language to build monolingual speech recognition systems. First, we investigate, whether in this case, mismatched transcription can be used together with matched transcription and still be helpful to improve performance. We show that PT adaptation can be improved further if we apply data augmentation for the matched training data. Second, we introduce a method to use both matched and mismatched transcriptions simultaneously in a multi-task learning framework. Finally, we investigate a two-level adaptation method that uses PT adaptation to generate alignment for multi-task learning.

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 describes the multi-task learning framework. Section 4 presents experimental setup. Experimental results are shown in Section 5. Section 6 concludes the paper.

2. Mismatched Transcription for Speech Recognition

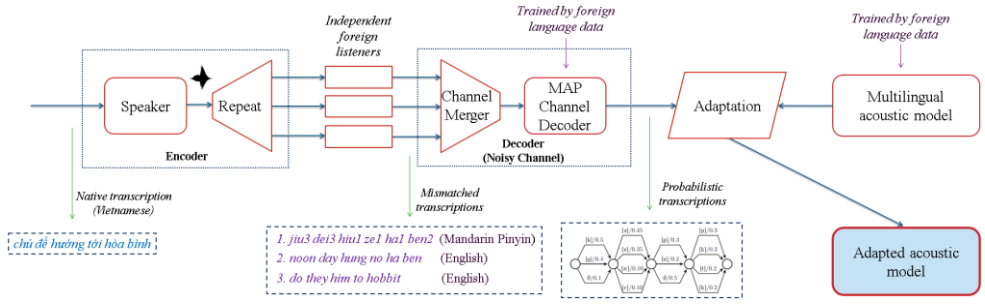


Figure 1: Mismatched transcription for speech recognition.

Mismatched crowdsourcing was recently proposed to solve the shortage of native transcription in some languages (Jyothi and Hasegawa-Johnson., 2015a; Jyothi and Hasegawa-Johnson., 2015b). Figure 1 illustrates the process of generating mismatched transcription and how to use it to improve speech recognition. A native speaker of the target language (under-resourced language) generates speech in his native language. Independent foreign transcribers of resource-rich annotation languages (crowd-workers) listen and write down nonsense syllables in the orthography of the annotation language (mismatched transcription). The non-native listeners are modeled as a communication channel that is "mismatched" to the message, in the sense that the channel (the listeners) uses an alphabet that is different from the alphabet of the message (the spoken utterance). A MAP channel decoder (https://github.com/uiuc-sst/PTgen) recovers, from observed mismatched transcriptions, a probability distribution over spoken phone strings, which we call a probabilistic transcription, and which is usually represented in the form of a phone lattice

(Hasegawa-Johnson et al., 2017). The probabilistic transcription can be used to adapt an existing acoustic model. In (Do et al., 2016), the authors further used a well-resourced ASR system to generate mismatched transcription, and found that combining mismatched transcription generated by both human and ASR leads to lower phone error rate.

3. Use Matched and Mismatched Transcriptions in a Multi-Task Learning Framework

One disadvantage of the approach in Figure 1 is that performance is reliant on the quality of the MAP channel decoder used to convert mismatched transcription to probabilistic transcription of the target language. The MAP channel decoder is only trained using limited parallel training data, i.e., audio with both matched and mismatched transcription; in the case of under-resourced languages, the channel model can be under-trained and information can be lost through this process.

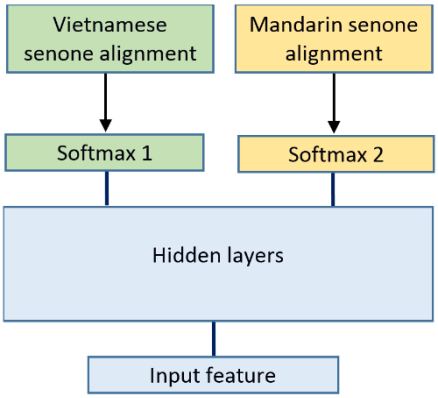


Figure 2: Multi-task learning DNN framework using both matched and mismatched transcription.

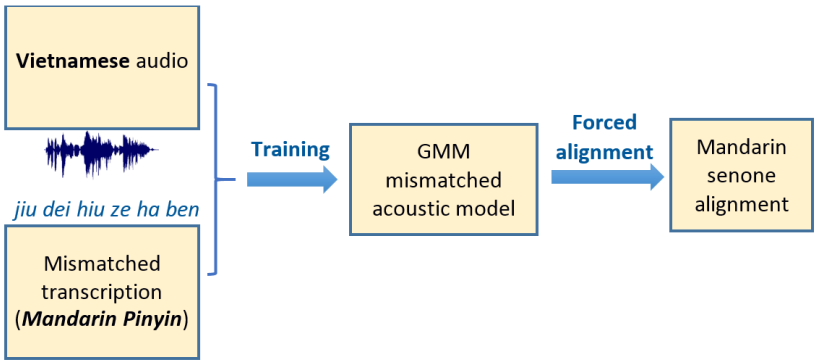


Figure 3: Target language (Vietnamese) audio and mismatched transcription (Mandarin Pinyin) are used to build the mismatched acoustic model.

In this paper, we investigate a method that uses mismatched transcription directly in a multi-task learning deep neural network (MTL-DNN) (Do et al., 2017). As shown in Figure 2, a MTL-DNN acoustic model has two softmax output layers, one for matched (target language - Vietnamese) transcription and one for mismatched (source language - Mandarin) transcription. Vietnamese frame alignments are generated through forced alignment using the initial Vietnamese GMM trained with limited Vietnamese data as in the conventional DNN training procedure or from the Vietnamese GMM model after applying PT adaptation (Liu et al., 2016). To obtain frame alignment for the mismatched transcription, we use a GMM mismatched acoustic model trained using the target language (Vietnamese) audio data with source language (Mandarin) mismatched transcription (Figure 3). The mismatched GMM acoustic model is then used to do forced alignment on the adaptation set to achieve frame alignment for DNN training. After training, the MTL-DNN can model speech perception of native listeners and foreign listeners simultaneously. This provides a natural structure to share speech perceptual characteristics of source and target language listeners on the target language speech.

Following our recent study (Do et al., 2017), cross-entropy is used as the objective function to train the MTL-DNN. Two cross-entropy functions are used for the two softmax layers, they are defined as follows.

Eq (1) for softmax 1 (matched: target language senones representing target language speech, e.g., Vietnamese senones trained on Vietnamese speech):

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

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 target language triphone state (senone).

Eq (2) for softmax 2 (mismatched: source language senones representing target language speech, e.g., Mandarin senones trained on Vietnamese speech):

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

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 matched GMM acoustic model, and k is a source language triphone state (senone).

The MTL-DNN is trained to minimize the following regularized multi-task objective function.

$$J = (1-\beta)p_2 J_1 + \beta p_1 J_2 \qquad (3)$$

where p_1, p_2 are the priors of training data size for matched and mismatched training data,

respectively, introduced to deal with data imbalance between two datasets. β is a tunable combination weight. When β=0, the MTL-DNN becomes a conventional DNN using only one target language softmax layer; when β=1, the MTL-DNN becomes a mismatched ASR, trained using audio in one language with transcriptions in another.

After DNN training, the softmax layer for mismatched transcription is discarded. Only the softmax layer for matched transcription (target language) is kept for decoding as in the conventional single-task DNN.

4. Experimental setup

In our experiments, Vietnamese is chosen as the under-resourced language and Mandarin speakers are chosen as non-native transcribers. The IARPA BABEL Vietnamese corpus provided in the context of the 2013 NIST Open Keyword Search Evaluation is used for our experiments. The acoustic data were collected from various real noisy scenes and telephony conditions. We randomly select 12, 24 and 48 minutes from the full training set with native transcription to simulate limited transcribed training data conditions. Together, 10 hours of untranscribed data are also selected for mismatched transcription. A total of 4 Mandarin speakers from Upwork (www.upwork.com) were hired, each in charge of 2.5 hours. Those Mandarin speakers listened to short Vietnamese speech segments and, for each segment, wrote a transcription in Pinyin alphabet that is acoustically closest to what they think they heard (Chen et al., 2016).

To convert mismatched transcription to probabilistic transcription, the MAP channel decoder is modeled as a finite memory process using a weighted finite state transducer (WFST). The channel decoder accepts phone sequences of the foreign language (Mandarin) and produces a Vietnamese phone lattice. The weights on the arcs of the WFST are learned using the EM algorithm (Dempster et al., 1977) to maximize the likelihood of the observed training instances using, as training data, audio for which both matched and mismatched transcriptions exist (three different WFSTs are trained, using 12, 24, or 48 minutes of parallel matched/mismatched transcriptions). The USC/ISI Carmel finite-state toolkit[1] is used for EM training of the WFST model and the OpenFST toolkit (Allauzen et al., 2007) is used for all finite-state operations. The Kaldi speech recognition toolkit (Povey et al., 2011) is used to build the GMM and DNN acoustic models.

A feature vector including 23 log-filterbank features and 3 pitch features is extracted every 10 milliseconds, using a 25-millisecond analysis window. Acoustic models are GMM with speaker adaptive training (SAT) and DNN. During the decoding process, a bigram

[1] *"Carmel finite-state toolkit," http://www.isi.edu/licensedsw/carmel/*

phonetic language model trained from training data is used. Performance of each system is evaluated using phone error rate (PER) on 20 minutes extracted from the 10-hour development set specified in the IARPA BABEL Vietnamese corpus. In this study tones are not considered; all tonal marks are removed.

5. Experimental Results

In this section, we present the results of using mismatched transcription by applying probabilistic transcription (PT) adaptation and multi-task learning on the initial models. Finally, we show the results by combining both PT adaptation and multi-task learning.

5.1. Probabilistic Transcription Adaptation

We first investigate performance of Vietnamese phone recognition when very limited transcribed training data are available. As shown in the first row of Table 1, PER of Vietnamese phone recognizer trained with 12 minutes of transcribed data is 83.98% for the GMM and 83.76% for the DNN systems. The main reasons for these high PERs are: the corpus is noisy conversational telephone speech and the training set is only 12 minutes.

#	Data augmentation	#State	w/o PT adaptation (initial model)		w/ PT adaptation (adapted model)	
			GMM	DNN	GMM	DNN
1	No	200	83.98	83.76	83.29	82.04
2	Yes	152	81.34	82.03	80.92	80.81
3	Yes	201	81.25	81.97	80.75	80.62
4	Yes	288	81.12	81.79	80.34	80.51
5	Yes	435	81.37	81.92	80.02	80.12
6	Yes	658	81.59	82.23	79.88	79.82
7	Yes	897	82.21	82.49	80.09	79.93
8	Yes	1450	82.94	83.01	80.56	80.35

Table 1: Phone Error Rate (PER %) for different acoustic models with 12 minutes of original Vietnamese matched training data.

To investigate the usefulness of mismatched transcription, two adaptation approaches are conducted using PT adaptation method (Section 2):

- GMM is adapted using MAP adaptation (Liu et al., 2016).
- DNN is adapted by further training with mismatched transcription (Das and Hasegawa-Johnson., 2016).

The last two columns of the first row in Table I represent the PER of the two above adaptation approaches, they are 83.29% and 82.04% for the adapted GMM and DNN models, respectively. We can see that using PT adaptation improves both the GMM and DNN models. However this improvement is not large. Our hypothesis is that because the number of triphone states in the acoustic model is relatively small, i.e., 200, the model is not improved effectively from a large amount of mismatched transcription i.e., 10 hours. However, we cannot easily increase the number of triphone states in the acoustic model since only 12 minutes of matched transcription are available. To overcome this, the data augmentation approach is considered. It is a common strategy adopted to increase the data quantity to avoid overfitting and improve the robustness of the model against different test conditions (Jaitly et al., 2013). In this study, we increase the training data size using a data augmentation technique called audio speed perturbation (Ko et al., 2015). Speed perturbation produces a warped time signal, for example, given speech waveform signal $x(t)$, time warping by a factor α will generate signal $x(\alpha t)$. In our experiment, we generate 3 copies of the original speech data with α={0.9, 1.0, 1.1}. Now, with a 3 times bigger data size, we can vary the model complexity by changing the number of triphone states from 152 to 1450.

Row 2 to row 8 of Table 1 illustrate performance of different models when data augmentation is applied. By comparing with row 1, we can see that data augmentation can help to improve the performance of Vietnamese phone recognizers significantly. The GMM and DNN models without using PT adaptation (initial models) achieve the best performance when the number of triphone states is 288. However, both the GMM and DNN using PT adaptation obtain the best performance when the number of triphone states reaches 658. This proves that our hypothesis is correct: to achieve the best effect of PT adaptation, the complexity of the initial acoustic model should be increased even beyond the point at which performance of the initial acoustic model is optimal. This can be explained as follows: when the number of triphone states increases, the model parameters cannot be well estimated using matched data alone, due to lack of training data of some triphones. This makes performance of the model drop. However, when a large amount of mismatched transcription is used to adapt this initial model, we have enough data to estimate model parameters of all the triphone states. One way to quantify the information provided by mismatched crowdsourcing is to count the number of distinct triphone labels in the transcription. In 12 minutes of matched transcription, there are 2506 distinct triphones. By examining the 1-best path through the probabilistic transcription i.e., the MAP estimate of the target language transcription after decoding the mismatched transcriptions, we found that there are 738 new triphones introduced which do not exist in the original matched training transcription. Figure 4 illustrates that if we have more matched transcription, the number of new triphones

introduced by the mismatched transcription reduces gradually. It proves that with sufficient matched training data, it's possible to find most of the new triphones generated by mismatched transcription. Hence, in the case of limited matched transcription, mismatched transcription can provide more statistics about triphones that would be otherwise unobserved.

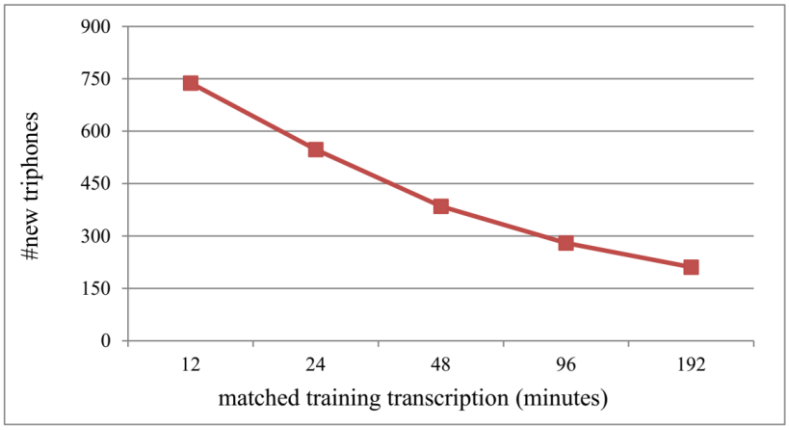


Figure 4: Number of new triphones introduced by the mismatched transcription versus amounts of matched training transcription.

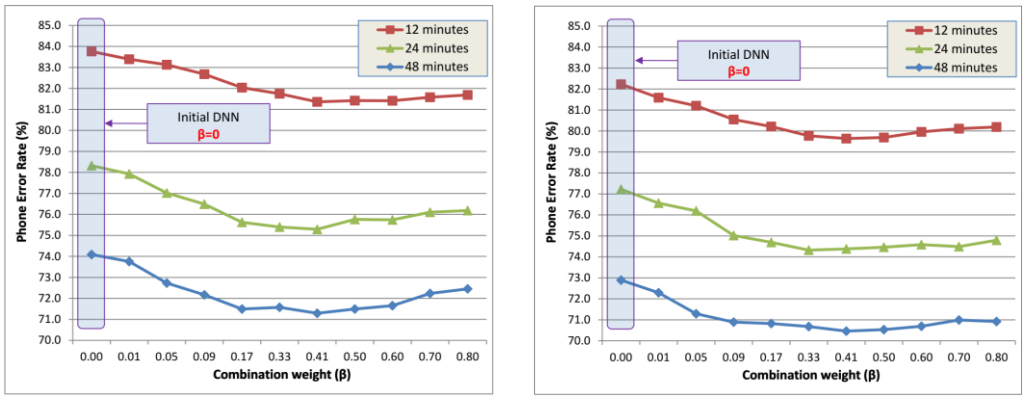
#	Data size	Data augmentation	Without PT adaptation (initial model)		With PT adaptation (adapted model)		Multi-task learning (β=0.41)	
			GMM (ALI1)	DNN	GMM (ALI2)	DNN	ALI1	ALI2
			(1)	(2)	(3)	(4)	(5)	(6)
1	12 minutes	No	83.98	83.76	83.29	82.04	81.36	80.34
2		Yes	81.59	82.23	79.88	79.82	79.64	78.52
3	24 minutes	No	78.51	78.32	77.94	76.82	75.29	73.74
4		Yes	76.92	77.22	75.83	75.60	74.38	73.02
5	48 minutes	No	75.05	74.09	74.38	73.28	71.29	70.60
6		Yes	73.49	72.89	72.77	71.92	70.46	70.04

Table 2: Phone Error Rate (PER %) for different acoustic models with different amounts of matched training transcription.

Our above experiments have shown that using PT adaptation together with data augmentation can significantly improve performance of Vietnamese phone recognition when 12 minutes of matched transcription are available. In the next experiments, we will investigate whether we still gain from mismatched transcription and data augmentation when more matched transcription is available. We conduct experiments with 24 and 48 minutes of matched transcription training data. As shown in Table 2, using data augmentation with PT adaptation achieves a consistent improvement for both the GMM and DNN models even with more training data, i.e., 24 and 48 minutes.

5.2 Multi-task Learning

Figure 5.a shows PER given by the MTL framework (Figure 2) for the case of 12, 24 and 48 minutes of matched transcription. The combination weight, β is varied from 0 to 0.8. $\beta = 0$ is the case of conventional initial DNN with only one matched data softmax layer as shown in the second column of Table 2. We can see that when β increases, the MTL framework can consistently improve performance for all three cases. When $\beta = 0.41$, we achieve the best performance with 81.36%, 75.29%, 71.29% 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.



(a) without data augmentation (b) with data augmentation

Figure 5: Phone error rate versus combination weight β in the multi-task learning framework.

Figure 5.b shows a similar observation when MTL is applied after Vietnamese speech data are augmented using speech perturbation. The PERs given by the MTL at $\beta = 0.41$, both without and with using data augmentation, are presented in column 5 of Table 2. It can be

seen that MTL significantly outperforms both the GMM and DNN with and without PT adaptation.

In the experiments whose results are shown in Fig. 5, frame alignment for the matched output layer of the DNN is provided by the initial GMM trained with limited matched training data (i.e., ALI1, the first column of Table 2). Table 2 also shows that by using PT adaptation for the GMM (ALI2), we obtain consistent improvement. Our previous study indicated that MTL can be significantly improved by using better frame alignment (Do et al., 2017). In this study, we use PT-adapted GMM to generate frame alignment (ALI2) for MTL. This can be considered as two level-adaptation where the first adaptation level is for the GMM, to generate better alignments, and the second level is for the DNN, to improve acoustic modeling. The last column of Table 2 is the PER given by MTL using frame alignment ALI2. In this case the combination weight β is simply set to 0.41. It shows that combining MTL with PT adaptation results in a consistent improvement over MTL in column 5 and PT adaptation in column 4.

6. Conclusion

In this paper, we presented the results of using mismatched transcription to improve performance of speech recognition of an under-resourced language. Experiments conducted on the IARPA BABEL Vietnamese corpus showed that mismatched transcription can significantly improve performance when matched transcription is limited. We also showed that using data augmentation for the matched training data makes mismatched transcription adaptation more effective. Finally, multi-task learning, a method using mismatched transcription directly, can be effectively combined with PT adaptation in order to build a two-level adaptation framework.

7. References

Allauzen, C., Riley, M., Schalkwyk, J., Skut, W., and Mohri, M., 2007, OpenFST: A general and efficient weighted finite state transducer library, *Implementation and Application of Automata*, pp. 11-23.

Burget, L., et al., 2010, Multilingual acoustic modeling for speech recognition based on subspace Gaussian mixture models, in Proc. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4334-4337.

Chen, W., Hasegawa-Johnson M., and N. F. Chen, 2016, Mismatched crowdsourcing based language perception for under-resourced languages, *Procedia Computer Science*, vol. 81, pp. 23–29.

Das, A., and Hasegawa-Johnson, M., 2016, An investigation on training deep neural

networks using probabilistic transcriptions, in Proc. *Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 3858-3862.

Dempster, A. P., Laird, N. M., and Rubin, D. B., 1977, Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society*, Series B 39(1), pp. 1-38.

Do, V. H., Chen, N. F., Lim, B. P., and Hasegawa-Johnson, M., 2016, Analysis of Mismatched Transcriptions Generated by Humans and Machines for Under-Resourced Languages, in Proc. *Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 3863-3867.

Do, V. H., Xiao, X., Chng, E. S., and Li, H., 2013, Context dependent phone mapping for LVCSR of under-resourced languages, in Proc. *Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 500–504.

Do, V. H., Xiao, X., Chng, E. S., and Li, H., 2014a, Cross-lingual phone mapping for large vocabulary speech recognition of under-resourced languages, *the IEICE Transactions on Information and Systems*, Vol. E97-D, No. 2, pp. 285–295.

Do, V. H., Xiao, X., Chng, E. S., and Li, H., 2014b, Kernel Density based Acoustic Model with Cross-lingual Bottleneck Features for Resource Limited LVCSR, in Proc. *Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 6–10.

Do, V. H., Chen, N. F., Lim, B. P., and Hasegawa-Johnson, M., 2017, Multi-Task Learning using Mismatched Transcription for Under-Resourced Speech Recognition, in Proc. *Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 734–738.

Hasegawa-Johnson., M., et al., 2017, ASR for Under Resourced Languages from Probabilistic Transcription, *IEEE/ACM Transaction on Audio, Speech and Language*, vol. 25, no. 1, pp. 46–59.

Imseng, D., Bourlard, H., and Garner, P. N., 2012, Using KL divergence and multilingual information to improve ASR for under-resourced languages, in Proc. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4869–4872.

Jyothi, P., and Hasegawa-Johnson, M., 2015a, *Acquiring speech transcriptions using mismatched crowdsourcing*, in Proc. AAAI, pp. 1263-1269.

Jyothi, P., and Hasegawa-Johnson, M., 2015b, Transcribing continuous speech using mismatched crowdsourcing, in Proc. *Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 2774-2778.

Jaitly, N., and Hinton, G. E., 2013, Vocal tract length perturbation (VTLP) improves speech recognition, in Proc. *ICML, Workshop on Deep Learning for Audio, Speech, and Language*

Processing, pp. 625-660.

Ko, T., Peddinti, V., Povey, D., and Khudanpur, S., 2015, Audio augmentation for speech recognition, in Proc. *Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 3586-3589.

Liu, C., et al., 2016, Adapting ASR for under-resourced languages using mismatched transcriptions, in Proc. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5840-5844.

Povey, D., et al., 2011, The Kaldi speech recognition toolkit, in Proc. *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*.

Sainath, T. N., et al., 2012, Exemplar-based processing for speech recognition: An overview, *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 98-113.

Schultz, T., and Waibel, A., 2001, Experiments On Cross Language Acoustic Modeling, in Proc. *International Conference on Spoken Language Processing (ICSLP)*, pp. 2721-2724.

Sim, K. C., and Li, H., 2008, Context Sensitive Probabilistic Phone Mapping Model for Cross-lingual Speech Recognition, in Proc. *Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 2715-2718.

Stolcke, A., Grezl, F., Hwang, M., Lei, X., Morgan, N., and Vergyri, D., 2006, Cross-domain and cross-language portability of acoustic features estimated by multilayer perceptrons, in Proc. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 321-324.

Vu, N. T., Kraus, F., and Schultz, T., 2011, Cross-language bootstrapping based on completely unsupervised training using multilingual A-stabil, in Proc. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5000–5003.

Xu, H., Do, V. H., Xiao, X., and Chng, E. S., 2015, A Comparative Study of BNF and DNN Multilingual Training on Cross-lingual Low-resource Speech Recognition, in Proc. *Annual Conference of the International Speech Communication Association (INTERSPEECH)*.