Analysis of Mismatched Transcriptions Generated by Humans and Machines for Under-Resourced Languages

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

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

vanhai.do@adsc.com.sg, {nfychen, bplim}@i2r.a-star.edu.sg, jhasegaw@illinois.edu

Abstract

When speech data with native transcriptions are scarce in an under-resourced language, automatic speech recognition (ASR) must be trained using other methods. Semi-supervised learning first labels the speech using ASR from other languages, then re-trains the ASR using the generated labels. Mismatched crowdsourcing asks crowd-workers unfamiliar with the language to transcribe it. In this paper, self-training and mismatched crowdsourcing are compared under exactly matched conditions. Specifically, speech data of the target language are decoded by the source language ASR systems into source language phone/word sequences. We find that (1) human mismatched crowdsourcing and cross-lingual ASR have similar error patterns, but different specific errors. (2) These two sources of information can be usefully combined in order to train a better target-language ASR. (3) The differences between the error patterns of non-native human listeners and non-native ASR are small, but when differences are observed, they provide information about the relationship between the phoneme systems of the annotator/source language (Mandarin) and the target language (Vietnamese).

Index Terms: speech recognition, semi-supervised learning, mismatched crowdsourcing, under-resourced languages

1. Introduction

Among the several thousands of spoken languages on Earth, only a few of them have been studied by the speech recognition community [1]. One of the main hurdles of Automatic Speech Recognition (ASR) system deployment in new languages is that an ASR system relies on a large amount of labeled training data for acoustic modeling. This makes a full-fledged acoustic modeling process impractical for under-resourced languages. To deal with this issue, several approaches have been proposed. The first approach is to transfer well-trained acoustic models to under-resourced languages, e.g., using a universal phone set [2, 3], tandem acoustic features [4–6], subspace GMMs (SGMMs) [7, 8], Kullback-Leibler divergence HMM (KL-HMM) [9, 10], and cross-lingual phone mapping [11–15]. The second approach attempts to increase the amount of labeled training data through active learning [16, 17] or semi-supervised learning [18, 19]. Recently, mismatched crowdsourcing was proposed as a potential approach to deal with the lack of native transcribers to produce labeled training data [20, 21]. In this method, the transcribers do not speak the under-resourced language of interest, yet, they write down what they hear in this language into nonsense words in their native language. The mismatched transcriptions are then decoded by a mismatched channel implemented by weighted finite state transducers. The experimental results in [20, 21] showed that using mismatched transcriptions improves performance of speech recognition systems over the multilingual or semi-supervised training approaches.

In this paper, semi-supervised learning techniques and mismatched crowdsourcing are compared under exactly equivalent conditions. A semi-supervised learner is constructed by training an ASR in a resource-rich language, and applying it to transcribe unlabeled data in the language of interest. The detailed error patterns of the cross-lingual ASR are compared to those of human crowd workers with exactly the same language background, i.e., native speakers of the same resource-rich language. With this approach, we can quickly generate large amounts of mismatched transcriptions without the need of hiring transcribers. To evaluate the quality of such mismatched transcriptions, we propose a normalized entropy index as a quality indicator.

This study also analyzes human and machine mismatched transcriptions. We show that mismatched transcriptions generated by humans and machines exhibit confusion patterns that are similar yet different in interesting ways: while the general trends are strikingly similar, the differences suggest that the ASR system chooses to use phonetic boundaries that are different from humans, yet these phonetic boundaries are phonologically meaningful from the perspective of second language acquisition [22]. In addition, we also observe that further improvement can be achieved by combining these two types of mismatched transcriptions.

2. Methods

In this section, we first introduce mismatched transcriptions and their applications for under-resourced language ASR. After that normalized entropy is proposed to use as a quality indicator for foreign ASR. Finally, the combination of human and machine mismatched transcriptions is presented.

2.1. Mismatched transcriptions

Mismatched crowdsourcing was recently proposed to solve the shortage of native transcription in some languages [20, 21]. As shown in Figure 1, the input to the system is a message, X, in the under-resourced utterance language, which is implemented

as a speech signal S. Nonnative transcribers (speakers of a resource-rich annotation language) listen to S, and write nonsense syllables, Y, in the orthography of the annotation language; Y is called the mismatched transcription. A decoder is used to estimate X given Y.

Decoding can be done using the maximum likelihood rule [21].

$$\tilde{X} = \text{argmax}\, p(X|Y) = \text{argmax}\, p(Y|X)\, p(X) \quad (1)$$

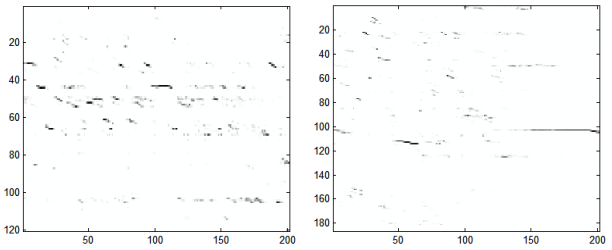
This process is similar to the conventional decoding process in ASR in which Y are the input features, X is the text, p(Y|X) computed by the acoustic model while p(X) is computed by the language model.

Figure 1: *Mismatched transcriptions for speech recognition: the target language is Vietnamese, the foreign language is English.*

In this paper, instead of using nonnative transcribers, machines (ASR systems trained in a resource-rich language) are used to generate mismatched transcriptions Y, from speech S. ASR systems well trained from the source language can generate good quality mismatched transcriptions for speech of the target language.

2.2. Use normalized entropy as a quality indicator

If several cross-lingual ASR systems are available to generate mismatched transcriptions, then it is useful to have a criterion for choosing the system likely to best transcribe speech in a target language. Phonetic overlap between the source and target languages must be considered, but other factors are also important, e.g., acoustic model architecture, corpus recording condition, speaking style, and corpus size.



(a) English (b) Hungarian

Figure 2: *Phoneme posteriorgrams of a Vietnamese segment given by English and Hungarian phoneme recognizers, x-axis is time in frame, y-axis is phoneme-state ID.*

While with human transcribers, it is intractable to obtain the confidence score for each utterance/word/phoneme, it is simple with ASR. In this paper, we evaluate quality of a foreign ASR system based on a type of confidence score called posterior probabilities. Posterior probability $p(q_i|o_t)$ provided by an acoustic model is probability of a speech class q_i (e.g., phoneme-state) given the input frame observation o_t. Figure 2 illustrates two phoneme posteriorgrams of the same Vietnamese speech segment provided by two phoneme recognizers, English and Hungarian [25]. Our hypothesis is that if the posteriorgram is sharp and clear that means the ASR system can clearly distinguish acoustic units in the target language and vice versa. In this case, the Hungarian recognizer is clearly a better choice for Vietnamese speech data. For each frame o_t, posterior probability $p(q_i|o_t)$ satisfies the constraint:

$$\sum_{i=1}^{N} p(q_i|o_t) = 1 \quad (2)$$

Where N is number of speech classes in the ASR such as number of phoneme-states. Hence for each frame o_t, posterior probabilities form a categorical distribution $\{p(q_i|o_t)\}$ and we can use frame-based entropy to estimate the sharpness of the distribution.

$$H_t = -\sum_{i=1}^{N} p(q_i|o_t)\log(p(q_i|o_t)) \quad (3)$$

To evaluate the quality of an ASR system for a speech corpus, we can use the average entropy by computing frame-based entropy H_t of all frames in the development set.

$$\bar{H} = \frac{1}{T}\sum_{t=1}^{T} H_t \quad (4)$$

where T is number of frames in the development set.

However, the number of speech classes N in different ASR systems can be different, hence the dynamic range of the average entropy varies over different ASR systems from 0 to log(N). In this paper, we use normalized entropy H_{norm} to evaluate the quality an ASR system.

$$H_{norm} = \frac{\bar{H} - \min(\bar{H})}{\max(\bar{H}) - \min(\bar{H})} = \frac{\bar{H}}{\log(N)} \quad (5)$$

2.3. Combination of human and machine mismatched transcriptions

Suppose we wish to train ASR in the target language by combining human and machine-based mismatched transcriptions. The simplest way to combine mismatched transcriptions is to treat the machine as a human transcriber, and apply the channel merging technique developed for mismatched crowdsourcing [21]. Though they have similar error rates, however, the mismatched transcriptions generated by humans and machines differ in some details, e.g., length of the transcription, therefore simple combination is suboptimal. To solve this, we transform machine mismatched transcription to be more human-like before doing combination, as shown in Figure 3. In this initial work, the converter is implemented as a WFST. The WFST is trained using the EM algorithm [23].

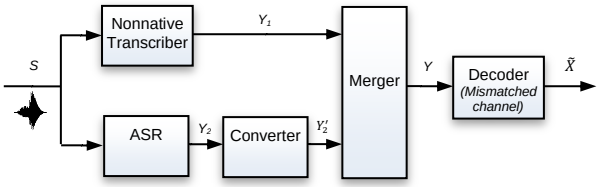


Figure 3: *Combination of mismatched transcriptions generated by humans and ASR.*

3. Experiments

3.1. Experimental setup

In our experiments, Vietnamese is chosen as the under-resourced language. The Vietnamese speech corpus was downloaded from the Australian Special Broadcasting Service consisting of mostly spontaneous, semi-formal speech (*http://www.sbs.com.au/podcasts/yourlanguage/vietnamese*). Bumpers and non-speech audio were discarded. There are 50 minutes of Vietnamese speech data, in which 40 minutes are used for training and 10 minutes are used to evaluate the performance. In this study tones are not considered; all tonal marks are removed.

To achieve human mismatched transcriptions, two sets of crowd workers are used: 10 English speakers from Amazon Mechanical Turk and 3 Mandarin speakers from Upwork. Each crowd worker listens to a short Vietnamese speech segment and writes down a transcription that is acoustically closest to what they think they heard [24]. For English and Mandarin speakers the mismatched transcriptions are in the form of English words and Pinyin alphabet, respectively. Native Vietnamese speakers were also recruited to provide native transcriptions.

To achieve machine mismatched transcriptions, different foreign ASR systems are used. First, we use 4 phoneme recognizers from the Brno University of Technology (BUT) [25]: Czech, Hungarian, Russian and English. Second, we use different ASR systems developed at the Institute for Infocomm Research (I^2R) for English and Mandarin. These systems were trained with 900, 2700 speech hours for the English and Mandarin acoustic models, respectively. Note that in this study, only the first recognition hypothesis (1-best) is used with machine mismatched transcriptions.

To convert mismatched transcriptions to matched transcriptions, a mismatched channel is used and modeled as a finite memory process using WFST. The input of the channel is phoneme sequences of the foreign language while the output is Vietnamese phoneme sequences. The weights on the arcs of the WFST model are learned using the EM algorithm [23] to maximize the likelihood of the observed training instances. The USC/ISI Carmel finite-state toolkit [26] is used for EM training of the WFST model and the OpenFST toolkit [27] is used for all finite-state operations. During the decoding process, unigram phonetic language models trained from training data are used.

3.2. Human vs. machine mismatched transcriptions

Table 1 shows the target Vietnamese language phoneme error rate (PER) of different systems. The left part of the table is the results from human mismatched transcriptions. The right part is the results when BUT recognizers are used to generate mismatched transcriptions. The last row of Table 1 represents the normalized entropy (H_{norm}) for different phoneme recognizers. We have three observations.

Table 1. *Phoneme error rate (PER) and normalized entropy (H_{norm}) for human and machine mismatched transcriptions. ENG: English, CMN: Mandarin, HUN: Hungarian, CES: Czech, RUS: Russian.*

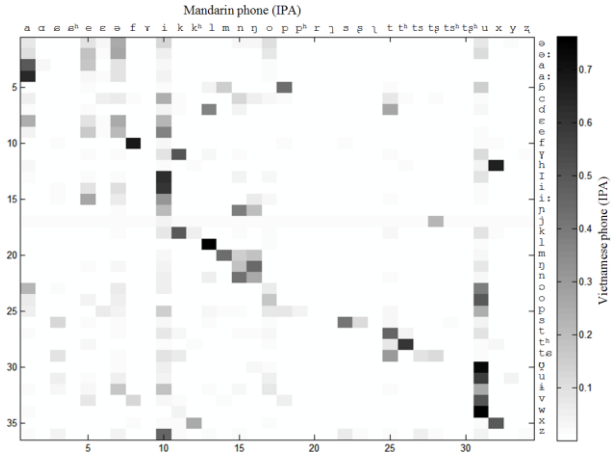
	Human mismatched transcription		Machine mismatched transcription			
	ENG	CMN	ENG	HUN	CES	RUS
PER (%)	76.02	69.20	97.00	75.42	75.69	84.70
H_{norm}	-	-	0.250	0.166	0.180	0.219

Among human mismatched transcriptions, Mandarin transcribers give better performance for Vietnamese than English transcribers. We speculate that Mandarin transcribers may be able to transcribe Vietnamese more accurately than English transcribers because the syllable structures of Vietnamese resemble Mandarin more than English [24].

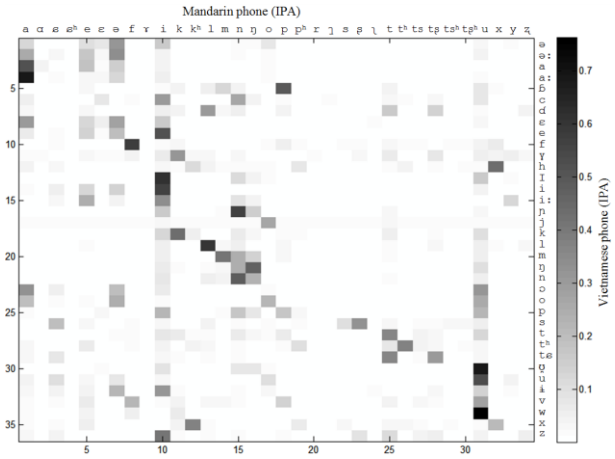
Second, for machine mismatched transcriptions, Czech and Hungarian recognizers provide similar performance, where these results are slightly better than the result given by English human transcribers. The English recognizer gives a much poorer result than other recognizers because it was trained using only 3 hours of speech data from the TIMIT corpus. The other three recognizers were trained with much more data in the SpeechDat-E corpus [28].

Third, as shown in the last row of Table 1, normalized entropy (H_{norm}) has a high correlation to the PER (corr=0.976).

In the next step, we use I^2R ASR systems to generate mismatched transcriptions. First, the Mandarin ASR with a free syllable loop is used, which results in 78.37% PER in Vietnamese. Second, two English ASR systems using free phone loop and free word loop (37k word vocabulary) are used which provide 71.74% and 84.41% PER in Vietnamese, respectively. We can see that to generate mismatched transcription using a free phone loop recognizer without any language constraints in the foreign language is a better choice. These results are also much better than the results provided by the TIMIT-based English system in Table 1, because they were trained using far more data. It can be concluded that using better foreign ASR systems can provide better mismatched transcriptions for the target language.



(a) human



(b) machine

Figure 4: *Weight matrices of the WFST mismatched channel for the language pair Mandarin-Vietnamese.*

Now, we make a deeper comparison between mismatched transcriptions generated by humans and machines. We focus on the language pair Mandarin-Vietnamese. Figure 4 shows

two weight matrices of the WFST mismatched channel, one for human and one for machine mismatched transcriptions. The x-axis represents the source language (Mandarin) phonemes; the y-axis represents the target language (Vietnamese) phonemes. Both matrices share similar patterns. However, there are several significant differences. Figure 5 shows the weights of the human and machine WFSTs for the same phoneme /s/ in Vietnamese. We can see that human speakers of Mandarin classify it one way, whereas Mandarin ASR classifies it a different way. One hypothesis compatible with this observation is as follows: the Vietnamese /s/ has acoustic characteristics between those of the Mandarin /s/ and the Mandarin /ʂ/. In support of this hypothesis, Fig. 6 shows periodogram spectral estimates (average of squared magnitude FFTs from several consecutive 6ms windows) for one arbitrarily-chosen example of each fricative.

To conclude, we find that humans and machines generate similar error patterns, but different specific errors. The two sources of information can be usefully combined in order to train a better target-language ASR. The differences between the error patterns of non-native human listeners and non-native ASR are small, but when differences are observed, they seem to provide information about the relationship between the phoneme systems of the annotator language (Mandarin) and the utterance language (Vietnamese).

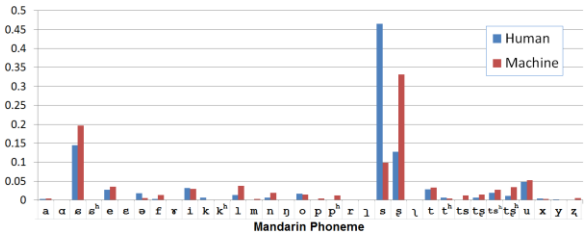


Figure 5: *The weights of the WFST mismatched channel for phoneme /s/ in Vietnamese.*

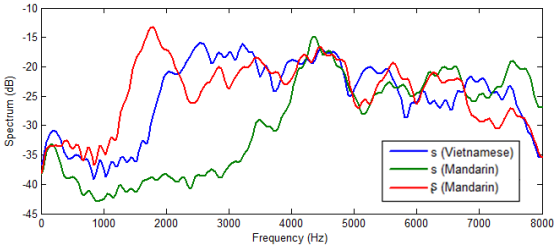


Figure 6: *Average spectra of phoneme /s/ and /ʂ/ in Vietnamese and Mandarin.*

3.3. Combination of human and machine mismatched transcriptions

Table 2. *PER for individual and combined systems for the language pair, Mandarin-Vietnamese.*

Individual system		Combined system	
Human	Machine	w/o conversion	w/ conversion
69.20	78.37	67.73	66.40

The simplest way to combine mismatch transcriptions generated by humans and machines is to treat machine as a human transcriber, and now we have one more input stream. As shown in the third column of Table 2, with this setup, we achieve 67.73% PER for the combined system which is 1.5% better than the human system (69.20%) and 10.6% better than the machine system (78.37%).

Despite similar error rates, mismatched transcriptions generated by humans and machines implicitly use a different encoding system to transcribe the acoustic observations. To account for such differences, machine mismatched transcription is converted to human-like transcription before doing combination, as shown in Figure 3. In this work, the converter is implemented as a WFST. In this case, both the input and output of the converter are Mandarin transcriptions. Hence the weights of the WFST converter can be drawn as a square matrix in Figure 7 where both the x-axis and y-axis represent Mandarin phonemes. We can see a lot of confusions between similar phonemes recognized by humans and machines such as /ɛ/ and /ɛʰ/; /k/ and /kʰ/; /n/ and /ŋ/; /s/ and /ʂ/; /ts/ and /tʂ/. After conversion, the converted machine transcriptions are merged with human mismatched transcriptions before mapping to the Vietnamese target language. As shown in the last column of Table 2, with this setup, we obtain PER of 66.40% which is slightly better than the 67.73% achieved using raw machine transcriptions for combination. For the future work, we will investigate more efficient ways to combine mismatched transcriptions generated by humans and machines.

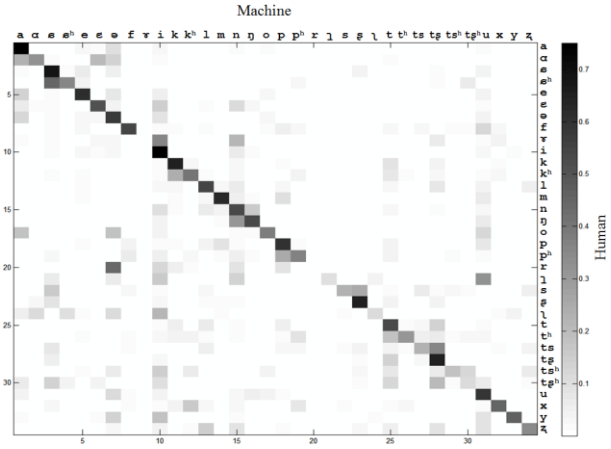


Figure 7: *Weight matrix of the WFST converter from machine to human transcriptions.*

4. Conclusions

This paper presented an alternative approach to achieve mismatched transcriptions using ASR systems of foreign languages. With this approach, we are able to generate large amounts of mismatched transcriptions without the need of hiring transcribers. To effectively evaluate the quality of the foreign ASR system for the target language, we proposed normalized entropy as a quality index. Experiments showed that by using mismatched transcriptions generated by machine can achieve a similar performance as human. In addition, the normalized entropy has been shown as a good quality index since it has a high correlation to the phoneme error rate of the target language. We also investigated the differences between mismatched transcriptions generated by humans and machines which led to an improvement by combining them.

5. Acknowledgements

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).

6. References

[1] H. Li, K. A. Lee, and B. Ma, "Spoken Language Recognition: From Fundamentals to Practice," *Proceedings of the IEEE*, Vol. 101, No. 5, May 2013, pp. 1136–1159.

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

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

[4] A. Stolcke, F. Grezl, M. Hwang, X. Lei, N. Morgan, and D. Vergyri, "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)*, 2006, pp. 321-324.

[5] S. Thomas, S. Ganapathy, and H. Hermansky, "Cross-lingual and multistream posterior features for low resource LVCSR systems," in Proc. *Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2010, pp. 877-880.

[6] P. Lal, "Cross-lingual Automatic Speech Recognition using Tandem Features," Ph.D. thesis, The University of Edinburgh, 2011.

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

[8] L. Lu, A. Ghoshal, and S. Renals, "Maximum a posteriori adaptation of subspace Gaussian mixture models for cross-lingual speech recognition," in Proc. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012, pp. 4877–4880.

[9] D. Imseng, H. Bourlard, and P. N. Garner, "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)*, 2012, pp. 4869–4872.

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

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

[12] K. C. Sim and H. Li, "Stream-based Context-sensitive Phone Mapping for Cross-lingual Speech Recognition," in Proc. *Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2009, pp. 3019-3022.

[13] K. C. Sim, "Discriminative Product-of-expert Acoustic Mapping for Crosslingual Phone Recognition," in Proc. *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2009, pp. 546-551.

[14] V. H. Do, X. Xiao, E. S. Chng, and H. Li, "Context dependant phone mapping for cross-lingual acoustic modeling," in Proc. *International Symposium on Chinese Spoken Language Processing (ISCSLP)*, 2012, pp. 16–20.

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

[16] G. Riccardi, and D. Hakkani-Tür, "Active learning: Theory and applications to automatic speech recognition," *IEEE Transactions on Speech and Audio Processing*, 13(4), 2005, pp. 504-511.

[17] D. Hakkani-Tur, G. Riccardi, and A. Gorin, "Active learning for automatic speech recognition," in Proc. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2002, pp. 3904-3907.

[18] D. Yu, B. Varadarajan, L. Deng, and A. Acero, "Active learning and semi-supervised learning for speech recognition: A unified framework using the global entropy reduction maximization criterion," *Computer Speech & Language*, 24(3), 2010, pp. 433-444.

[19] S. Thomas, et al. "Deep neural network features and semi-supervised training for low resource speech recognition)," in Proc. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 6704-6708.

[20] P. Jyothi and M. Hasegawa-Johnson, "Acquiring speech transcriptions using mismatched crowdsourcing," in Proc. *AAAI*, 2015.

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

[22] P. Escudero, "Linguistic perception and second language acquisition: Explaining the attainment of optimal phonological categorization," PhD thesis, Netherlands Graduate School of Linguistics, 2005.

[23] 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 39(1), 1977, pp. 1–38.

[24] W. Chen, M. Hasegawa-Johnson and N. F. Chen, "Mismatched crowdsourcing based language perception for under-resourced languages," in Proc. *International Workshop on Spoken Language Technologies for Under-resourced Languages (SLTU)*, 2016.

[25] "Phoneme recognizer based on long temporal context," http://speech.fit.vutbr.cz/software/phoneme-recognizer-based-long-temporal-context

[26] "Carmel finite-state toolkit," http://www.isi.edu/licensed-sw/carmel/

[27] C. Allauzen, M. Riley, J. Schalkwyk, W. Skut, and M. Mohri, "OpenFst: A general and efficient weighted finite-state transducer library," in Proc. *the Ninth International Conference on Implementation and Application of Automata (CIAA)*, 2007.

[28] D. Caseiro, I. Trancoso, "Spoken language identification using the SpeechDat corpus", in Proc. *International Conference on Spoken Language Processing (ICSLP)*, 1998, pp. 3197-3200.