# RECOGNIZING ZERO-RESOURCED LANGUAGES BASED ON MISMATCHED MACHINE TRANSCRIPTIONS

*Wenda Chen*[1,2,*], *Mark Hasegawa-Johnson*[1], *Nancy F. Chen*[2]

[1]Beckman Institute, UIUC, USA
[2]Institute for Infocomm Research, A*STAR, Singapore

## ABSTRACT

Mismatched crowdsourcing based probabilistic human transcription has been proposed recently for training and adapting acoustic models for zero-resourced languages where we do not have any native transcriptions. This paper describes a machine transcription based phone recognition system for recognizing zero-resourced languages and compares it with baseline systems of MAP adaptation and semi-supervised self training. With a set of available speech recognizers in source languages that cover all the basic phonetic features, this work shows that we can use mismatched machine transcriptions from these source languages to achieve human level transcriptions, bypassing the laborious efforts of obtaining human transcriptions. We also present a fully automated unsupervised approach for zero-resourced speech recognition using mismatched machine transcriptions for transfer learning of phone models.

***Index Terms***— automatic speech recognition (ASR), mismatched machine transcription, zero-resourced languages, modular system, transfer learning

## 1. INTRODUCTION

For recognizing zero-resourced languages where the native transcriptions are missing, we had previous work on generating probabilistic transcriptions (PT) from mismatched crowdsourcing based human non-sense transcriptions. The transcribers don't understand the target language and are transcribing the target speech using English or Mandarin orthographies. The PTs are then used for training or adapting a multilingual ASR system [1, 2, 3].

For the task of generating transcription labels for training ASR, could we create a fully automated machine transcription system that reads the speech signals using English and Mandarin recognizers and achieves better results than the human mismatched transcripts which are sometimes very noisy? Given the clustering approach in [4] and the distinctive feature knowledge, could we replace the crowdsourcing based human transcriptions with machine transcriptions and design a corresponding speech recognition system? We would like to address these issues in the paper. Vietnamese and Singapore Hokkien (Hokkien,[5]) are the zero-resourced languages we are using to analyze the usefulness of machine transcription systems. We do not use any native transcriptions in the experiments.

## 2. RELATED WORK

Mismatched transcripts have been proven to be useful in acoustic modeling for speech recognition on zero-resourced languages. The transcribers are presented with audio clips of the unknown target language and asked to use the orthography of their native language to write down what they hear. The resulting syllabic words from the transcriptions are then converted and interpreted as phone level probabilistic transcriptions (PT), which can be used for acoustic modeling [1, 2, 6].

With mismatched transcripts available in two languages (e.g., Mandarin and English transcriptions of Vietnamese), we recently showed that improved probabilistic transcripts are obtained by clustering the alignments between the annotator languages [7]. It represents the alignments in a bipartite graph based matrix where each entry represents the probability of phone mappings. The clusters are then obtained iteratively from the matrix to simulate the process of extracting the closest phone clusters that the annotators from different language backgrounds used to represent the same target speech. We then proposed an optimization framework for inferring clusters of the phonemes or graphemes in two annotators' languages that were used to represent the closely related phonemes in the target language [4]. The resulting phonetic clusters also automatically represent the interaction between tone and phone, similar to the tone-dependent phone sets of ASR.

When we have available resources to train acoustic models of a related rich-resourced language, we could also use these data for a zero-resourced target languages. Various approaches have been proposed, such as bootstrapping from source-model alignments [8, 9],pooling data across languages [10], adaptation and self training of the neural network models [11, 12], and phone mapping for recognition with the source models [13]. We are still lacking the clear procedures on the best way to perform such cross-lingual sharing and need to evaluate the benefits that can be expected for different amount of source and target data.

This paper will first show the extend to which we can use fully automatic system to replace human crowdsourcing system. It is the first time we propose a systematic way of deciding and using mismatched machine transcriptions for zero-resourced language speech recognition. It then proposes an unsupervised cluster based phone recognition system followed by a language model. Phone recognition followed by language modeling (PRLM) has been successfully adopted in other speech processing fields such as language recognition [14] and similar concepts have also applied to spoken language summarization [15]. We will show the effectiveness of PRLM system in the task of speech recognition for zero-resourced languages.

## 3. PROPOSED USAGE OF MACHINE TRANSCRIPTIONS

We can transcribe speech into English and Mandarin phones simultaneously using speech recognition systems and then cluster them into target phone sequence. By considering both the acoustic patterns

[*]This work is funded by the Agency for Science, Technology and Research (A*STAR) Graduate Scholarship.

and linguistic knowledge, it could potentially improve the current human probabilistic transcription based system.

### 3.1. Steps of Machine Mismatched Transcription Algorithm

**Step 1**. Recognize the target speech using English and Mandarin phone recognizers, such as the BUT phone recognizers[16] and $I^2R$ speech recognizers[17], respectively. We collect the word level outputs from multiple available word recognizers, such as Google, CMU Sphinx, BUT, $I^2R$, as the different machine transcribers and then convert them to phone level sequences using lexicons. The results will be compared based on recognition languages to find the more generalizable one to better recognize the target language. In this paper, recognition in a set of languages (Hungarian, English, Mandarin, Czech, and Russian) are selected and used to generate a set of phone error rates for comparison.

**Step 2**. Align the Mandarin and English (or other selected languages) phone sequences using Minimum Edit Distance based on distinctive features from linguistic knowledge of the languages [6, 18] and then derive the clusters using the clustering process as in [4]. This makes use of the distinctive feature knowledge to characterize the phone differences between the languages [19]. It provides additional information to the acoustic models.

**Step 3**. Convert the aligned phone recognition results from the multiple recognizers to cluster sequences and use the majority vote method to determine the final recognition results at target phone level based on the clustering mapping derived in step 2. Evaluate the phone error rate of the predicted transcripts.

### 3.2. Clustering Algorithm

The clustering algorithm proposed earlier [4] first constructs a matrix W that represents a bipartite graph. Each entry value in W is defined as

$$w_{ij} = \frac{1}{N} \sum_{q \in \mathcal{X}_i} S_q(j)$$

where $S_q(j)$ is the substitution probability for Mandarin phoneme j by English transcription token $q$, $\mathcal{X}_i$ is the set of all transcription instances of the $i^{\text{th}}$ English grapheme, and $N$ is the number of all transcription segments in the training data. The normalized distance between any subsets A and B of English and Mandarin phonemes is defined as

$$d_N(A,B) = \frac{d(A,B)}{W(A,M)+W(E,B)} + \frac{d(A^c,B^c)}{W(A^c,M)+W(E,B^c)}.$$

where

$$d(A,B) = W(A,B^c) + W(A^c,B)$$

and

$$W(A,B) = \sum_{i \epsilon A, j \epsilon B} w_{ij}$$

It is shown that minimizing $d_N(A,B)$ is equivalent to finding the second largest singular vectors of the matrix $D_X^{-\frac{1}{2}} W D_Y^{-\frac{1}{2}}$, where $D_X$ and $D_Y$ are the diagonal matrices where each diagonal element is the sum of the corresponding row or column of W. The resulted sub-clusters A and B are to be divided again according to the algorithm. Eventually the phones in each cluster are closely related and can be tagged with one target phone in the target language based on phonetic feature similarity.

### 3.3. Language Set Analysis

| | |
|---|---|
| **English Trans.** | DH EY AA B EH N EY UW G<br>AA K Y NG AH M OW OP AY V<br>EY B AW W IH DH EH AH |
| **Mandarin Trans.** | HUA2 BEN4 SHU3 GANG1 HAO3<br>JUN2 AN1 MAO4 BAI3 ZEI2<br>LENG3 DA3 WEI3 BIE2 |

**Table 1**. *Sample utterance in Vietnamese with mismatched machine transcriptions in English phones and Mandarin Pinyin with 4 tones.*

Sample utterance in Vietnamese with mismatched machine transcriptions at phone level in English and Mandarin Pinyin are shown in Table 1. It shows that automatic speech recognizers can detect the pronunciations of consonants and vowels in the speech signals. In our experiments, 2 transcribers or ASR systems (from BUT, $I^2R$ and CMU Sphinx) are used for English and Mandarin. One system is used for other languages. The phone error rates of different systems and the clustering results from two language pairs (English+Mandarin and Hungarian+Mandarin) are presented in Table 2. We can observe that, due to the different language similarities based on the distinctive features, certain mismatched languages that have similar phoneme pronunciations can be used to transcribe the similar languages in a better way (English+Mandarin performs worse than Hungarian+Mandarin).

The cost or difficulty of getting transcribers from certain language is related to the number of available transcribers or translators of certain language we can find in Upwork (www.upwork.com). For example, we can find 159683 transcribers of English, 3746 transcribers in Mandarin and 1286 transcribers in Hungarian. Then it is much more costly to find Hungarian transcribers than finding English transcribers[18].

| **PER of recognizers** | **Vietnam.** | **Hokkien** |
|---|---|---|
| **Hungarian** | 74.97% | 73.42% |
| **Mandarin** | 78.37% | 72.51% |
| **English** | 84.41% | 83.20% |
| **Czech** | 75.69% | 74.56% |
| **Russian** | 84.70% | 87.70% |
| **Cluster(English+Mandarin)** | 76.32% | 71.31% |
| **Cluster(Hungarian+English)** | 75.94% | 72.59% |
| **Cluster(Russian+Mandarin)** | 79.86% | 74.64% |
| **Cluster(Hungarian+Russian)** | 78.15% | 75.32% |
| **Cluster(Mandarin+Czech)** | 76.93% | 69.13% |
| **Cluster(Hungarian+Czech)** | 74.62% | 70.56% |
| **Cluster(Hungarian+Mandarin)** | 74.11% | 67.42% |

**Table 2**. *Phone Error Rate (PER) of different recognition systems from BUT (Hungarian, English, Czech, Russian), $I^2R$ (English and Mandarin), and CMU Sphinx (Mandarin)*

For a given target language, we can find in the language coverage table [18] on the closest transcribing languages that cover all the distinctive features in the target language. Since each language has a weight which indicates the number of transcribers/translators in Upwork, we have to choose the transcribers and native languages according to the difficulty and cost of getting them. For our example as in Table 2, Hungarian+Mandarin turns out to be the best combination for both Vietnamese and Hokkien. The English+Mandarin transcription of Vietnamese is less accurate than the corresponding Hungarian+Mandarin combination. This indicates that if possible, getting Hungarian transcribers will be performing better than getting English transcribers. However, since Hungarian transcribers are difficult to be found online, we may better use mismatched machine transcriptions in Hungarian and Mandarin in this case in stead of using human English+Mandarin transcriptions online.

According to the language coverage table for the distinctive features, the languages Hungarian, Mandarin, English, Czech and Russian could theoretically cover all the distinctive features of most languages. Hence for any zero-resourced language, we could test it using the recognition systems in these five languages. If any one of the systems give a phone recognition accuracy higher than a threshold, we would suggest to use machine transcriptions and then the proposed modular system, instead of using human transcriptions. Next, we will propose such a system and try to find such a threshold.

| PER of transcripts | Vietnam. | Hokkien |
|---|---|---|
| **PT English** | 76.02% | 70.34% |
| **Clustering(Human)** | 68.45% | 67.96% |
| **Clustering(Machine)** | 74.11% | 67.42% |

**Table 3**. *Phone Error Rate from predicted transcriptions. PT English is the probabilistic transcript from English transcribers. Clustering(Human) is the clustering of English and Mandarin transcriptions. Cluster(Machine) is the clustering of the outputs of Hungarian and Mandarin recognizers.*

Clustering method makes it possible to compare the machine and human transcriptions on the same settings of predicted transcriptions. The results from the proposed machine transcription framework are compared with the results from the previous probabilistic transcription approach in Table 3. Mismatched transcribers from English background typically provide non-sense word transcriptions that are more noisy for grapheme-to-phoneme conversion, compared with the Mandarin transcriptions [20], due to the nature of the languages. In Table 3, we can observe that for Vietnamese, since human Mandarin transcribers do a good job on transcription, the resulted human transcriptions are better than the machine transcriptions. For Hokkien, the machine transcriptions from Hungarian and Mandarin outperforms the human transcriptions from English and Mandarin. It shows that the performance of the existing speech recognizers on the target language does provide reasonable links to predict whether the machine transcriptions are comparable with the human transcriptions when we evaluate on the cluster sequence.

## 4. UNSUPERVISED MODULAR PHONE RECOGNITION SYSTEMS USING MISMATCHED CLUSTERS

This section discusses the usage of automatically generated clusters for proposing and developing an unsupervised phone recognition system.

### 4.1. Use clusters to combine the English and Mandarin mismatched recognizers

For a typical Neural Network (NN) structure, we have input x at each aligned time frame to be the acoustic features and output to the corresponding clusters derived from the previous section. Assume that in each column of the weight matrix W, we define the cluster which corresponds to the target segment to be

$$C_j = \{M_1, M_2, ..M_k\} \cup \{E_1, E_2, ..E_l\}$$

i.e. the cluster $C_j$ contains $k$ Mandarin phones and $l$ English phones. The weighted input at the output layer given the input x should be

$$Pr(C_j|x) = \sum_{k=1}^{N_M} Pr(M_k|x)W_M(k,j) + \sum_{l=1}^{N_E} Pr(E_l|x)W_E(l,j)$$

Output:

$$Y_{kc} = \frac{Pr(C_k|x_c)}{\sum_{j=1}^{V} Pr(C_j|x_c)}$$

where $N_M$ and $N_E$ are the total number of Mandarin and English phones and acoustic models, respectively, $Pr(M_k|x)$ and $Pr(E_l|x)$ are fixed posteriors and given by the HMM-DNN or HMM-GMM trained acoustic models of typical Mandarin and English recognizers, and c is the time frame index. The softmax output is compared with the reference target phone labels V predicted from the clusters [4] (dimension: target phone set × number of phone labels) using the cross entropy criteria

$$E = -\sum_c \sum_j V_{jc} log(Y_{jc})$$

Hence the update rule for $W_E$, and likewise for $W_M$ is

$$\frac{\partial E}{\partial W_E(l,k)} = -\sum_c V_{kc} Pr(E_l|x)(1 - Y_{kc})$$

The weights in neural network define the soft boundaries of the clusters and the resulted network can be used to combine the model outputs of English and Mandarin recognition systems and generate target phone sequences during testing.

### 4.2. Parsing recognized outputs using integrated language model

After we have developed the phone recognition system as in figure 1 and the previous section, where the clusters are used to train the soft weightings of the corresponding English and Mandarin acoustic models, the cluster based language model is then developed here and used to parse the output from the speech recognizer to further improve the phone recognition accuracy. This modular system is inspired from chapter 7 in [21]. It shows that with the English and Mandarin recognizers and mismatched machine transcriptions available, clustering can help prepare a large amount of transcription labels for acoustic model training, without the need of hiring native transcribers. In this case, cluster sequences, when converted to IPA sequences, are used as the generated transcriptions to combine the output of the English and Mandarin recognizers.

Then we trained the cluster based language model with previous recognized clusters and current mismatched phones as inputs, the finally predicted cluster as output. In particular, we use LSTM model with the vector of $[C^{-1}, M^{-1}, E^{-1}]$, $[C_x, M, E]$ as the two-time-step input, C as the output, where C is the generated target cluster sequence, E and M are the English and Mandarin mismatched grapheme sequence, $C^{-1}, M^{-1}, E^{-1}$ are the C, M, E at previous time step, $C_x$ is the currently recognized cluster from the adapted acoustic model given the current audio frame x. The general parsing model's likelihood is hence $p(C|C_x, E, M, C^{-1}, E^{-1}, M^{-1})$. With mismatched crowdsourced data, this model is trained and used to do further correction on the output of the recognition system, assuming that we only know the number of phones in the target language.

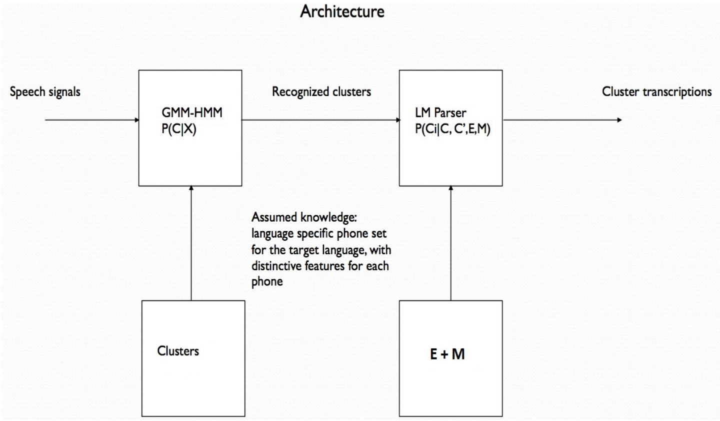

**Fig. 1**. *The phone recognition and language model (PRLM) modular system that combines sections 4.1 and 4.2*

## 5. EXPERIMENTAL RESULTS

### 5.1. Baseline 1 (Model based transfer learning): MAP adaptation system using PT

This paper proposed a database based transfer learning system. It will be compared with the model based transfer learning approaches for the target phones. In natural language processing research where the source data is labeled but the target data is unlabeled, it was presented that the database based transfer learning with domain adaptation is more effective than the model or feature based transfer learning [22]. Here the model transfer is performed with Maximum A Posteriori (MAP) adaptation using PT. In training the parameters of the baseline acoustic model in [2], for each training utterance, we work with the cascade $H \circ C \circ L \circ T$, where T is a linear chain FST representing the training transcript. During adaptation, for each training utterance (in the target language), we work with the cascade $H \circ C \circ L \circ PT$, where PT is a WFST representing the probabilistic transcript. During training of the acoustic models, we use the universal phone set and train the multilingual system using 40 minutes data from each of the six languages including Arabic, Dutch, Hungarian, Mandarin, Swahili and Urdu, in SBS dataset [23]. The experiment uses about 1 hour of mismatched transcriptions of Vietnamese (from SBS) and Hokkien (collected in $I^2R$ [4]) for adaptation and about 10 minute matched transcriptions for the evaluation.

### 5.2. Baseline 2 (Semi-supervised self training of acoustic models): Cluster-Trained Phone Recognition System

Self generated training labels are used to train acoustic models under semi-supervised settings with the best obtained labels [24]. This approach first learns the phone level bigram language model from the target phone sequences generated in section 3.1. Then it trains an ASR system using the target language's 1 hour speech data transcribed by humans and machines. It uses the best converted target phone cluster sequences as labels and the learned language model. Finally it uses the trained system as the phone recognizer to recognize 2 hr extra speech data in the target language to generate the self trained labels and then train the system again with the 3 hr speech data and the generated labels. It uses the same 10 minute matched transcriptions for evaluation.

### 5.3. Results: phone recognition

First, we train two monolingual DNN based ASR systems in English and Mandarin using Kaldi [25] and more than 40 hours of speech data for each language [17]. Our modular system then does not change the well trained English and Mandarin acoustic models themselves. Instead, we only combine the two monolingual systems using the soft clusters as an additional neural network layer and further parse the output using the learned context dependent relations from the cluster sequences. Hence it is not to re-train the system using the new labels but to transfer the knowledge in rich-resourced language to learn the zero-resourced language. We evaluated and compared the clustering based phone prediction performance generated from the transcripts by humans and machines in Table 3. In Table 4, we compare the phone recognition performances of the baseline MAP model transfer system, system self-trained using the predicted cluster labels and the modular system. Modular system further improves the performance of the phone recognition system, machine clustering system and model transfer baseline for both languages. Using rich-resource-trained speech recognizers in English and Mandarin and linguistic knowledge in distinctive features has reasonable benefits from the more accurate acoustic models compared with the system trained with limited 1 hour predicted cluster labels.

| PER of recognition systems | Vietnam. | Hokkien |
|---|---|---|
| **Baseline 1** | 76.61% | 72.78% |
| **Baseline 2** | 73.28% | 67.49% |
| **Proposed Modular System** | 69.17% | 66.54% |

**Table 4**. *Error rates from three phone recognition systems.*

## 6. DISCUSSION AND CONCLUSION

Proposed procedures for deciding human or machine mismatched transcriptions are related to the phone recognition accuracy of the existing speech recognition systems in a set of languages and the language coverage weightings. If the existing speech recognizers in certain language set can give an accuracy over a pre-defined threshold, we would suggest to use machine transcriptions. Here in our experiments, we find that the automatic phone recognition can use the machine transcriptions in a better way than the human transcriptions for Hokkien but not for Vietnamese. Since two speech recognition systems (Hungarian and Mandarin) can give a phone error rate less than 74% for Hokkien and none of them can have a similar result for Vietnamese, we can guess that such a threshold could be below 74% phone error rate. The exact threshold range and its generalizability need further studies.

Clustering method together with machine transcriptions are used as an automated mismatched phone recognition system. The phone recognizer followed by phone language model is then proposed. Machine mismatched transcriptions are comparable to human mismatched transcriptions for low-resourced ASR, given the constraint and trade-off that machine transcription can use any languages that better match the target language, while human transcription is limited to the resources of English and Mandarin transcribers (finding low-resourced language transcribers online is harder and more expensive according to the language coverage weight). When there is no rich-resourced language that is very close to the target language, and the machine transcriptions can give a performance higher than a proposed threshold, machine transcriptions are preferred for the zero-resourced languages that are hard to find native transcribers.

## 7. REFERENCES

[1] Preethi Jyothi and Mark Hasegawa-Johnson, "Acquiring speech transcriptions using mismatched crowdsourcing," *Proc. AAAI*, 2015.

[2] Chunxi Liu, Preethi Jyothi, Hao Tang, Vimal Manohar, Rose Sloan, Tyler Kekona, Mark Hasegawa-Johnson, and Sanjeev Khudanpur, "Adapting ASR for Under-Resourced Languages Using Mismatched Transcriptions," *Proc. ICASSP*, 2016.

[3] Mark Hasegawa-Johnson, Preethi Jyothi, Daniel McCloy, Majid Mirbagheri, Giovanni di Liberto, Amit Das, Bradley Ekin, Chunxi Liu, Vimal Manohar, Hao Tang, Edmund C. Lalor, Nancy Chen, Paul Hager, Tyler Kekona, Rose Sloan, and Adrian KC Lee, "Asr for under-resourced languages from probabilistic transcription," *IEEE/ACM Trans. Audio, Speech and Language*, vol. 25(1), pp. 46–59, 2017.

[4] Wenda Chen, Mark Hasegawa-Johnson, Nancy F Chen, and Boon Pang Lim, "Mismatched crowdsourcing from multiple annotator languages for recognizing zero-resourced languages: A nullspace clustering approach," *Proc. Interspeech*, 2017.

[5] Vanessa Lim, Hui Shan Ang, Estelle Lee, and Boon Pang Lim, "Towards an Interactive Voice Agent for Singapore Hokkien," *HAI '16 Proceedings of the Fourth International Conference on Human Agent Interaction*, pp. 249–252, 2016.

[6] Wenda Chen, Mark Hasegawa-Johnson, and Nancy F Chen, "Mismatched Crowdsourcing based Language Perception for Under-resourced Languages," *Procedia Computer Science*, vol. 81, pp. 23–29, 2016.

[7] Wenda Chen, Mark Hasegawa-Johnson, Nancy F. Chen, Preethi Jyothi, and Lav R. Varshney, "Clustering-based Phonetic Projection in Mismatched Crowdsourcing Channels for Low-resourced ASR," *6th Workshop on South and Southeast Asian NLP*, 2016.

[8] Schultz T. and Waibel A., "Language independent and language adaptive acoustic modeling for speech recognition," *Speech Communication*, vol. 35, pp. 3151, 2001.

[9] Le V.-B. and Besacier L., "Automatic speech recognition for underresourced languages: application to Vietnamese language," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 17(8), pp. 14711482, 2009.

[10] van Heerden C., Kleynhans N., Barnard E., and Davel M., "Pooling ASR data for closely related languages. in: Proceedings of the workshop on spoken languages technologies for under- resourced languages," *SLTU 2010, Penang, Malaysia*, p. 1723, May 2010.

[11] Frantisek Grzl, Martin Karafit, and Karel Vesely, "Adaptation of multilingual stacked bottle-neck neural network structure for new language," *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pp. 7654–7658, 2014.

[12] Karel Vesely, Mirko Hannemann, and Lukas Burget, "Semi-supervised training of deep neural networks," *Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on*, pp. 267–272.

[13] Chan H.Y. and Rosenfeld R., "Discriminative pronunciation learning for speech recognition for resource scarce languages," *In: Proceedings of the 2nd ACM Symposium on Computing for Development*, p. Article No. 12, 2012.

[14] M. A. Zissman, "Language identification using phoneme recognition and phonotactic language modeling," vol. 5, pp. 3503–3506 vol.5, May 1995.

[15] N. F. Chen, B. Ma, and H. Li, "Minimal-resource phonetic language models to summarize untranscribed speech," pp. 8357–8361, May 2013.

[16] P. Schwarz, "Phoneme recognition based on long temporal context, phd thesis," *Brno University of Technology*, 2009.

[17] Van Hai Do, Nancy F. Chen, Boon Pang Lim, and Mark Hasegawa-Johnson, "Analysis of mismatched transcriptions generated by humans and machines for under-resourced languages," *Interspeech*, 2016.

[18] Lav R. Varshney, Preethi Jyothi, and Mark Hasegawa-Johnson, "Language Coverage for Mismatched Crowd- sourcing," *Information Theory and Applications (ITA) Workshop, San Diego, California*, 2016.

[19] Steven Moran, Daniel McCloy, and Richard Wright [eds], "Phoible On Line," *Leipzig: Max Planck Institute for Evolutionary Anthropology (Available on line at http://phoible.org. Accessed on 2016-07-21)*.

[20] Van Hai Do, Nancy F. Chen, Boon Pang Lim, and Mark Hasegawa-Johnson, "Multi-task Learning for Phone Recognition of Under-resourced Languages using Mismatched Transcription," *IEEE/ACM Transactions on Audio, Speech and Language Processing, submitted*, 2017.

[21] Stephen E. Levinson, "Mathematical Models for Speech Technology," *John Wiley and Sons*, 2005.

[22] Sebastian Ruder and Barbara Plank, "Learning to select data for transfer learning with bayesian optimization," *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Copenhagen, Denmark*, p. 372382.

[23] Special Broadcasting Services Australia, "http://www.sbs.com.au/yourlanguage," .

[24] Scott Novotney and Richard Schwartz, "Analysis of low-resource acoustic model self-training," pp. 244–247, 01 2009.

[25] D. Povey and A. Ghoshal et. al., "The Kaldi Speech Recognition Toolkit," *ASRU*, 2011.