Mismatched Crowdsourcing From Multiple Annotator Languages For Recognizing Zero-resourced Languages: A Nullspace Clustering Approach

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

[1]Beckman Institute, University of Illinois at Urbana-Champaign, USA
[2] Institute for Infocomm Research, A*STAR, Singapore

{wchen113, jhasegaw}@illinois.edu, {nfychen,bplim}@i2r.a-star.edu.sg

Abstract

It is extremely challenging to create training labels for building acoustic models of zero-resourced languages, in which conventional resources required for model training – lexicons, transcribed audio, or in extreme cases even orthographic system or a viable phone set design for the language – are unavailable. Here, language mismatched transcripts, in which audio is transcribed in the orthographic system of a completely different language by possibly non-speakers of the target language may play a vital role. Such mismatched transcripts have recently been successfully obtained through crowdsourcing and shown to be beneficial to ASR performance. This paper further studies this problem of using mismatched crowdsourced transcripts in a tonal language for which we have no standard orthography, and in which we may not even know the phoneme inventory. It proposes methods to project the multilingual mismatched transcriptions of a tonal language to the target phone segments. The results tested on Cantonese and Singapore Hokkien have shown that the reconstructed phone sequences' accuracies have absolute increment of more than 3% from those of previously proposed monolingual probabilistic transcription methods.

Index Terms: mismatched crowdsourcing and perception, zero-resourced languages, automatic speech recognition

1. Introduction

The speech technology community has recently started considering the problem of automatic speech recognition (ASR) in unwritten languages [1, 2, 3, 4, 5, 6]. Linguists have created orthographic systems for half of the world's languages [11]; when we say that a language is "unwritten", therefore, what we really mean is that native speakers of the language do not know how to read and write their own language. Consider, for example, the cases of Cantonese, and of Singapore Hokkien. Cantonese is an official language in Hong Kong and Macau, spoken by 97 million people [7], most of whom receive primary education in Mandarin. Written Cantonese is standardized in Hong Kong, but most Cantonese speakers outside Hong Kong learn to read and write only in Mandarin [8]. Singapore Hokkien, a language related to Min Nan, is the native language of about 1.2 million people [9]. In Singapore, all native speakers of Hokkien under the age of 65 received primary education in Mandarin, English, Tamil or Malay. In the current literature, there are no official orthographies for the language. In Singapore Hokkien, place names and human names are written using a variety of similar but not identical conventions. Some historical orthographies for Min Nan or Hokkien spoken in China exist but have variation among them. Hong et al did phonological analysis of Singapore Hokkien based on fieldwork with three SH speakers [23]. In [16], Lim et. al. defined the Hokkien phones and trained the first Hokkien speech recognition system. We have obtained the transliteration and mismatched transcription of of 3.76 hours of Hokkien conversation data, which will be used for deriving the target phone clusters in Singapore Hokkien.

Speech recognition trained using mismatched crowdsourcing data has been proposed to be a useful tool for low resourced languages. The transcribers are presented with audio clips of the target language, which they usually don't understand, and use the orthography of their own languages to write down what they hear. The non-sense syllabic words from the transcriptions are then converted and interpreted as phone level probabilistic transcriptions (PT), which can be used to train automatic speech recognition [18,19, 22].

If mismatched transcripts are available in two languages (e.g., Mandarin and English transcriptions of Vietnamese), we recently showed that improved probabilistic transcripts are obtained by clustering alignments between the annotator languages [17]. It represents the alignments in a bipartite graph based matrix that 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 use to represent the target language. This paper proposes a new optimisation-based framework for inferring clusters of the graphemes in two annotator languages that are similar to the phonemes of the target language, where similarity is defined in terms of the probabilities of alignment between orthographic symbols. The resulting phonetic clusters automatically represent the interaction between tone and phone, in a representation similar to the tone-dependent phone sets of most ASR.This approach is different from other clustering algorithms in that it considers the total number of the phones in the target language, obtain the optimum number of representative clusters in the annotator language, and derives the weights for each cluster.

2. Prior Work: Zero-resourced speech recognition

Zero-resource speech recognition is generally defined as automatic speech recognition (ASR) trained using audio in the target language, but with no native language transcriptions. Several variations of this problem statement exist.

First, zero-resource spoken term discovery [13] involves clustering similar audio segments in a large untranscribed corpus. Frequently repeated long audio segments are taken to be keywords descriptive of the audio corpus.

If a language has no written text, the meaning of an utterance needs to be extracted in some form other than text. Harwath and Glass [14] proposed spoken term discovery using au-

Thanks to A*STAR Graduate Scholarship for funding.

dio captions of Flickr8k images; the meaning of the utterance is defined by correct retrieval of the desired image. Duong et al. [12] proposed translation from speech in an under-resourced language directly to text in a well-resourced language, without using text in the under-resourced language as an intermediate representation.

Acoustic unit discovery [1,5,6] differs from spoken term discovery in that it seeks to form clusters that account for all of the untranscribed speech. The clusters formed in this way ("acoustic units") are usually defined to be approximately the duration of a phonetic segment, thus the discovered units can be treated as an unsupervised approximation to the phone inventory of the target language. In Kamper's thesis, neural networks are applied on the acoustic features to obtain the higher level bayesian classification model [15].

A small number of studies have specifically explored the problem of discovering the phoneme inventory of an unwritten language [10]. The goal of phoneme discovery studies is not merely to discover acoustic units in the target language, but to associate each discovered unit with an IPA phone symbol, so that it is possible to generate an IPA phonetic transcription of speech in the target language without using any information about the (possibly non-existent) orthography of the target language.

3. Data Preparation and Description

The mismatched transcriptions of the target languages in English and Mandarin completed so far are described in Table 1. We used the Cantonese data from the Special Broadcasting Service Australia (SBS) audio corpus [26] and collected the Hokkien conversational speech spoken in Singapore. The Hokkien speech is then transcribed by local speakers using the proposed phone set and orthography [23]. Our multilingual mismatched crowdsourcing corpus consisted of one hour of Cantonese transcribed by 6 Mandarin transcribers and 10 English transcribers, and 3.76 hours of Hokkien language transcribed by by 2 English and 2 Mandarin transcribers per segment. Let's first examine the samples for Hokkien transcriptions in Table 2. From the transcriptions we can observe that the Mandarin Pinyin transcriptions and the English transcriptions simulate the Hokkien pronunciations in different ways. For example, Mandarin has 4 tones and Hokkien has 8 tones. The tone information in the Mandarin Pinyin transcription sometimes affects the transcriber's choice of a vowel. The Hokkien tones were not recorded by the native transcribers, because the task was too difficult for them.

Transcriptions	Audio	
Languages	Cantonese	Hokkien
Mandarin	48 (6)	221 (2)
English	68(10)	221(2)
Native	68(1)	26(1)

Table 1: *Summary of Transcription Data: Minutes of Audio in Each Language (Number of Transcribers who Annotated Each Audio Segment).*

Hokkien Native	u tsit pai ua ei lau pei to tia lahng gohng ah, i gohng hoh, ei, le tsi ku
English Trans.	oo che bai wei wei buh tee eh nuh koh wei kon oh eh lech go
Mandarin Trans.	wu3 qi1 bai4 wai4 wai4 lao3 bei3 dou1 tian1 lan2 gong1 ai3 gong1 luo2 ei1 le1 zi1 gou1

Table 2: *Sample utterance in Hokkien with mismatched transcriptions in English and Pinyin.*

4. Nullspace Clustering Algorithm

This section describes how we infer phone mappings between the transcriber languages and the target language. Specifically we describe a phonetic projection framework and clustering criteria with random projections. The problem is formulated as a bipartite graph clustering problem followed by segment classification. English grapheme labels and Mandarin grapheme labels are clustered based on their overlap in time (Figure 1 shows the algorithm). Phonological distinctive features of each cluster define the features of one inferred phone in the target language; accuracy of this process is evaluated by mapping the resulting pseudo-phone to the closest known phoneme in the Phoible phone inventory of the target language [24]. The experiment is evaluated with Cantonese and Singapore Hokkien.

4.1. Inferring a Phone Set Using Spectral Recursive Embedding

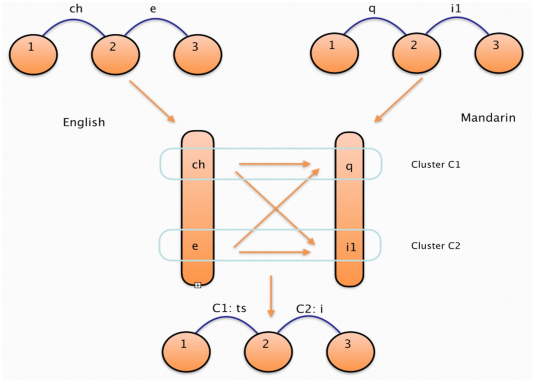

Figure 1: *Goal of the system: from English-Mandarin dynamic alignments to bipartite graph for clustering and constructing the predicted transcription labels. The sentence is from the real example in Table 2.*

Suppose that we have mismatched transcripts in Mandarin and English orthography, but we do not have native Cantonese phone transcripts. Additionally, let us assume that we do not know the Cantonese phone set. Take one of the two probabilistic transcripts (English, say) to define the number of Cantonese phone tokens per utterance. Align the other PT to it (the Mandarin one). The Mandarin PT has one or two orthographic symbols (or a deletion symbol) aligned to every segment of the English PT; thus for each English transcription token q, its substitution probability mass function (pmf) $S_q(j)$ has up to two nonzero entries, where j indexes a Mandarin grapheme type.

We first aggregate these probabilities over all tokens of the same English grapheme type, so that

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

where $\mathcal{X}_i$ is the set of all transcription instances of the i^{th} En-

glish grapheme, and N is the number of all transcription segments in the training data. Thus w_{ij} is the joint probability of Mandarin grapheme j and English grapheme i, i.e., the elements of the W matrix sum to one.

In order to avoid losing tone information, we define the Mandarin orthography to be composed of Pinyin onsets and tone-annotated rhymes. Thus, the sequence $< hai3, ya2, you1, len1 >$ is decomposed into the 8 graphemes $< h, ai3, y, a2, y, ou1, l, en1 >$, which are aligned to the English orthographic sequence $< ch, an, h, eihn, n, uw, l, ah >$.

After constructing the matrix W, we perform bipartite graph clustering using the spectral recursive estimation algorithm of [27]. To explain this algorithm, define E to be the set of all English graphemes, of which A is a subset; define M to be the set of all Mandarin graphemes, of which B is a subset. Generally, the similarity between two sets A and B where $A \subset E$ and $B \subset M$can be defined as:

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

Hence the distance $d(A,B)$ and normalized distance $d_N(A,B)$ between English grapheme set A and Mandarin grapheme set B can be computed using:

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

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

where c denotes set complement, $A \cup A^c = E$, $B \cup B^c = M$. The final optimization criterion is then $\min_{\pi(A,B)} d_N(A,B)$ where $\pi(A,B)$ denotes partitioning into the A and B clusters.

It is shown in [27] that

$$\min_{\pi(A,B)} d_N(A,B) = 1 - \max_{x \neq 0, y \neq 0} \{ \frac{2x^T W y}{x^T D_X x + y^T D_Y y} | x^T D_X e + y^T D_Y e = 0 \} \quad (1)$$

where e is the vector with all elements equal to 1, D_X and D_Y are the diagonal matrices where each diagonal element is the sum of the corresponding row or column of W, and $x = \{x_i\}$ and $y = \{y_i\}$ are the shifted version of the partitioning vectors $\bar{x} = \{\bar{x}_i\}$ and $\bar{y} = \{\bar{y}_i\}$ defined below.

Let $\bar{x}_i$ and $\bar{y}_j$ be the i^{th} and j^{th} elements of vectors $\bar{x}$ and $\bar{y}$, respectively; the clusters A and B are defined as

$$i \in A \text{ iff } \bar{x}_i > 0, \text{ else } i \in A^c \quad (2)$$

$$j \in B \text{ iff } \bar{y}_j > 0, \text{ else } j \in B^c \quad (3)$$

The largest left and right singular vectors of the matrix $D_X^{-1/2} W D_Y^{-1/2}$ are composed of uniformly non-negative elements, and are therefore not solutions of Equation (1). Equation (1) is therefore solved by the second largest left and right singular vectors ($\hat{x}$,$\hat{y}$) of the matrix $D_X^{-\frac{1}{2}} W D_Y^{-\frac{1}{2}}$. Then $x = D_X^{-\frac{1}{2}} \hat{x}$ and $y = D_Y^{-\frac{1}{2}} \hat{y}$.

The procedure in Equations (1) through (3) can be performed iteratively, developing a recursive bifurcation called Spectral Recursive Estimation [27]. Given a weighted bipartite graph with edge weight matrix W, we form partitions A for vertex set E, and B for vertex set M as the first cluster for the target segment. Subsequently we recursively partition the subgraphs $G(A,B)$ and $G(A^c,B^c)$ until we test and obtain the same number of clusters as the number of segments of the target language in Phoible. Each English and Mandarin grapheme in the clusters can be converted to IPA using the grapheme to phoneme mappings previously learned for English and Mandarin [17,19]. Then we know the distinctive features of the graphemes in each clusters. Hence for each distinctive feature in a cluster, we can compute the modal value of the distinctive features for all of the graphemes that make up the cluster. Once we have computed the modal distinctive feature vector for each cluster, we will choose the closest phoneme in the target language to be tagged with each cluster, by matching to the distinctive feature vectors in the Phoible inventory.

4.2. Validating the Clusters Using Null-Space Embedding

Theoretically, we define the relation between the target language segments and English segments as the projection $P_1 V = E$ where the V and E vectors are phone occurrence frequency in the target language matched transcriptions and grapheme occurrence frequency in English mismatched transcriptions, respectively. P_1 is a $dim(E) \times dim(V)$ matrix in which entry element (i,k) is the alignment frequency based probability that the k'th target phoneme is denoted by the i'th English grapheme. Hence we can compute the entropy of target language when transcribed in English from the P_1. In practice, we can only estimate V and P_1, for which purpose we define the following simplified estimate. We can consider each entry value of P_1, $P_1(i,k)$ as the probability that the corresponding target language phone k is correctly represented by English grapheme i. Based on the clustering result, we set $P_1(i,k) = 0$ if English grapheme i is not clustered to be in the cluster k in the target language. If the grapheme i is clustered into cluster k, then the entry (i,k) in P_1 is the normalized similarity between phone i and phone k estimated from the distinctive feature set in Phoible, i.e.

$$P_1(i,k) \propto \exp(-\frac{1}{n} \sum_{f=1}^{n} |i_f - k_f|).$$

where i_f is the value of distinctive feature f for grapheme i, e.g. in the Phoible inventory there are $n = 38$ distinctive features, and the constant of proportionality is chosen so that $\sum_i P_1(i,k) = 1$.. Likewise we define Mandarin as M and we have $P_2 V = M$.

Suppose we have computed (using, e.g., a machine translation toolkit) an alignment between English and Mandarin mismatched transcripts, and can compute the relation between English and Mandarin as $P_3 E = M$, where, in the notation of Section 4.1, $P_3 = W^T D_X$ is the matrix of conditional probabilities $P_3(j,i) = \Pr(j|i)$ of Mandarin grapheme j given English grapheme i. So we have $P_3 P_1 V = P_2 V$ which will lead to

$$(P_3 P_1 - P_2) V = 0. \quad (4)$$

In other words, V is in the null space of the projection matrix $P_3 P_1 - P_2$.

In our clustering algorithm, the clusters corresponding to the target segments are generated sequentially and tagged with the target segments based on the distinctive feature similarity. Hence we can obtain a vector V at each iteration of the cluster partitioning whose k^{th} value is 1 if the corresponding cluster is used to represent the target segment. Ideally, we want to map the segment dimensions of English and segment dimensions of

Mandarin with tones into the segment dimensions of target language. In practice, we may not need to use all the target segment clusters to achieve the optimum partitioning given the criteria in equation (4), since from information theory, we know that the entropy H(new segment clusters)≤H(full segments in the matched transcriptions of target language). Iterating over all of the phones in V, we find that Eq. (4) must be true even if V is a diagonal matrix, representing $\dim(V)$ consecutive applications of Eq. (4) to the $\dim(V)$ different phone-dependent subsets of the training data. If it were possible to perfectly distinguish all phonemes of the target language using mismatched transcriptions, it would be possible to set $V = I$, the identity matrix, and solve Equation (4) to find P_2 and P_1. In the non-ideal case, we initialize $V = I$, then allow the diagonal elements of V to shrink in the range $v_k \in [0, 1]$; small diagonal elements denote transcription segment clusters (target language phones) whose evidence from the mismatched transcripts is equivocal.

The null space clustering problem for finding the optimum V can be hence equivalently reformulated [25] as

$$V = argmin\frac{1}{2}||I - V||_F^2 + \frac{\lambda}{2}||(P_3P_1 - P_2)V||_F^2 \quad (5)$$

where V is a diagonal matrix, initialized as $V = I$ the identity matrix.The diagonal values could vary in the interval [0,1], representing the weight for each cluster. λ is tested to be 0.6 at optimum and $||\cdot||_F$ is Frobenius norm. Eq. (5) is convex in V, therefore optimising Eq. (5) results in a unique closed form solution derived in

$$(I + \lambda(WP_2 - P_1)^T(P_3P_2 - P_1))V = I$$

Eq. (5) is also balanced between a regulariser (the first term) that biases each column of V toward a delta-function centered on the corresponding target phone, and a penalty (the second term) that biases each column of V toward a solution in the nullspace of $P_3P_1 - P_2$. Upon convergence, the non-zeros in the k^{th} column of P_2V can be interpreted as the projection to Mandarin of the k^{th} target language phoneme, while the k'th column of P_1V is its projection to English; Eq. (5) finds the number of non-zero diagonal values of V as the number of phonemes in the target language that can be distinguished based on mismatched transcripts without causing a large difference between P_3P_1V and P_2V. The new number of clusters will further improve the accuracy of the predicted transcripts.

5. Experimental Results and Analysis

This section tests our clustering based method on Cantonese and Singapore Hokkien (SH), using matched data for evaluation, and compared with the simple majority vote method based on the alignment results, the mismatched channel based PT method and a general feed-forward Neural Network approach. Both methods are mappings from English and Mandarin graphemes to target phones trained from the aligned transcriptions of 10 minutes native language labels in the test corpus. To test on SH data, we collected 2 transcribers' transcriptions in each mismatched language and evaluated them by matched transcriptions. SH phone set mapping and conversion are performed according to [23] on the adaptation and evaluation sets. The sample predicted transcription output for the test SH utterance as in Table 2 is $< u, ts, b, ai, w, e, b, u, t, o, t, a, n, l, a, g, o, n, o, ei, l, e, g, o >$, which is clustered from the mismatched transcriptions in English and Pinyin. We observe the common phone alignment patterns of vowels such as 'oo' in English is often aligned to 'u3', 'u4' in Mandarin. As a result, 'oo', 'u3', 'u4' are eventually grouped in the same cluster and mapped to 'u' in SH. Similarly as another example for consonants, our algorithm successfully grouped 'ch' 'ts' in English and 'q' 'zh' in Mandarin to 'ts' in SH.

The first three rows in Table 3 are the lattice oracle results of the transcriptions of SH using Mandarin and English where we search the transcriptions to get the best results. The remaining seven rows show the results of multilingual PT and clustering algorithms applied on SH. The similar results for Cantonese were shown in Table 4. We observe the consistent improvement pattern on both Cantonese and SH's phone accuracies from monolingual PT, multilingual PT and then proposed clustering method. The optimisation procedures for the optimum number of cluster further improves the results from the clusters initially obtained from bipartite graph mining.

Probabilistic Transcription Source	Phone Error Rate
English	75.5%
Mandarin	69.3%
English+Mandariin (merged one-best)	64.1%
Majority Vote	87.1%
PT on English	76.6%
PT on Mandarin	73.4%
PT on E and M	70.3%
Neural Network	68.6%
Clustering method before optimization	67.5%
Clustering method after optimization	66.9%

Table 3: *The first three rows show the oracle lattice phone error rate (most correct path through the lattice) of the probabilistic transcriptions computed from three different sources. The next seven rows show the phone error rate (highest-scoring path through the lattice) of the SH probabilistic transcriptions computed from seven PT based and clustering based methods.*

Cantonese	Phone Error Rate
Majority Vote	65.1%
PT on English	64.3%
PT on Mandarin	47.4%
PT on E and M	43.1%
Neural Network	40.7%
Clustering method before optimization	39.1%
Clustering method after optimization	38.7%

Table 4: *Phone error rate (PER) for PT methods on Cantonese speech data. The oracle results were reported in [19].*

6. Conclusion and Future Work

This work demonstrates a new method for merging mismatched crowdsourcing from two different annotator languages. The proposed method finds phoneme clusters, each of which is similar to one phoneme of the target language, where similarity is quantified by the probability of alignments between orthographic symbols in the two different annotator languages (as given by the matrices P_1, P_2, and P_3). The results have shown the usefulness of the clustering method to improve the phone recognition of under-resourced languages. The future work will be modeling the acoustic units using the learned clusters for speech recognition without prior lexicon knowledge.

7. References

[1] Sujeeth Bharadwaj, "A Theory of (Almost) Zero Resource Speech Recognition." Ph.D. Thesis, University of Illinois, 2015

[2] Emmanuel Dupoux, Odette Scharenborg, and Graham Neubig, "Learning grounded linguistic units for languages without orthography," Proposal for a 2017 Jelinek Speech and Language Technology workshop.

[3] David Harwath, T.J. Hazen and J.R. Glass, "Zero resource spoken audio corpus analysis," ICASSP 2013, 8555-8559

[4] Aren Jansen, Ken Church and Hynek Hermansky, "Towards spoken term discovery at scale with zero resources," in Interspeech 2010, 1676-1679

[5] Lee and Glass, "A Nonparametric Bayesian Approach to Acoustic Model Discovery," 2012

[6] Lucas Ondel, Lukas Burget, and Jan Cernocky. Variational inference for acoustic unit discovery. In 5th Workshop on Spoken Language Technology for Under-resourced Language, 2016.

[7] Accredited Language Services, "Cantonese," 2017. Downloaded 3/11/2017 from https://www.alsintl.com/resources/languages/Cantonese.

[8] Donald B. Snow, Cantonese as Written Language: The Growth of a Written Chinese Vernacular. Hong Kong University Press, 2004.

[9] Adrien Tien, "Chinese Hokkien in Singapore: evidence for an indigenised Singapore culture." National University of Singapore, August 2013.

[10] Timothy Kempton and Roger K. Moore, Discovering the phoneme inventory of an unwritten language: A machine-assisted approach, jsc, pages 152 – 166, Elsevier, vol. 56, 2014

[11] Wycliffe Global Alliance, "Scripture and Language Statistics, 2016," downloaded 3/11/2017 from http://www.wycliffe.net/statistics.

[12] L Duong, A Anastasopoulos, D Chiang, S Bird, T Cohn, An attentional model for speech translation without transcription, Proceedings of NAACL-HLT, 949-959

[13] A. Park and J. Glass, Unsupervised Pattern Discovery in Speech, Trans. ASLP, 16(1), 186-197, 2008.

[14] David Harwath, James Glass, Deep multimodal semantic embeddings for speech and images, 2015/12/13, Automatic Speech Recognition and Understanding (ASRU), 2015 IEEE Workshop on, Pages 237-244

[15] Herman Kamper, Unsupervised neural and Bayesian models for zero-resource speech processing, CoRR, 2017

[16] Vanessa Lim, Hui Shan Ang, Estelle Lee, Boon Pang Lim, Towards an Interactive Voice Agent for Singapore Hokkien, HAI '16 Proceedings of the Fourth International Conference on Human Agent Interaction, Pages 249-252

[17] Wenda Chen, Mark Hasegawa-Johnson, Nancy F. Chen, Preethi Jyothi, Lav R. Varshney, Clustering-based Phonetic Projection in Mismatched Crowdsourcing Channels for Low-resourced ASR, 6th Workshop on South and Southeast Asian NLP, Dec 11, COLING 2016

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

[19] Wenda Chen, Mark Hasegawa-Johnson, and Nancy F Chen, Mismatched Crowdsourcing based Language Perception for Under-resourced Languages, Procedia Computer Science, Volume 81, 2016, Pages 23-29

[20] D. Povey, A. Ghoshal et. al, The Kaldi Speech Recognition Toolkit, ASRU 2011

[21] Jean-Luc Gauvain and Chin-Hui Lee, Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains, Speech and Audio Processing, IEEE Transactions on, Volume 2, 291–298, 1994

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

[23] Hong Yu Qing Amelia, A Phonological and phonetic Description of Singapore Hokkien, B. A. thesis, Nanyang Technological University, 2012

[24] 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. Access ed on 2016-07-21)

[25] P Ji, Y Zhong, H Li, M Salzmann, Null space clustering with applications to motion segmentation and face clustering, 2014 IEEE International Conference on Image Processing (ICIP), 283-287

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

[27] Hongyuan Zha, Xiaofeng He, Chris Ding, Horst Simon, and Ming Gu, Bipartite Graph Partitioning and Data Clustering, 2001, Proceedings of the tenth international conference on information and knowledge management (CIKM 2001), pages 25-32