

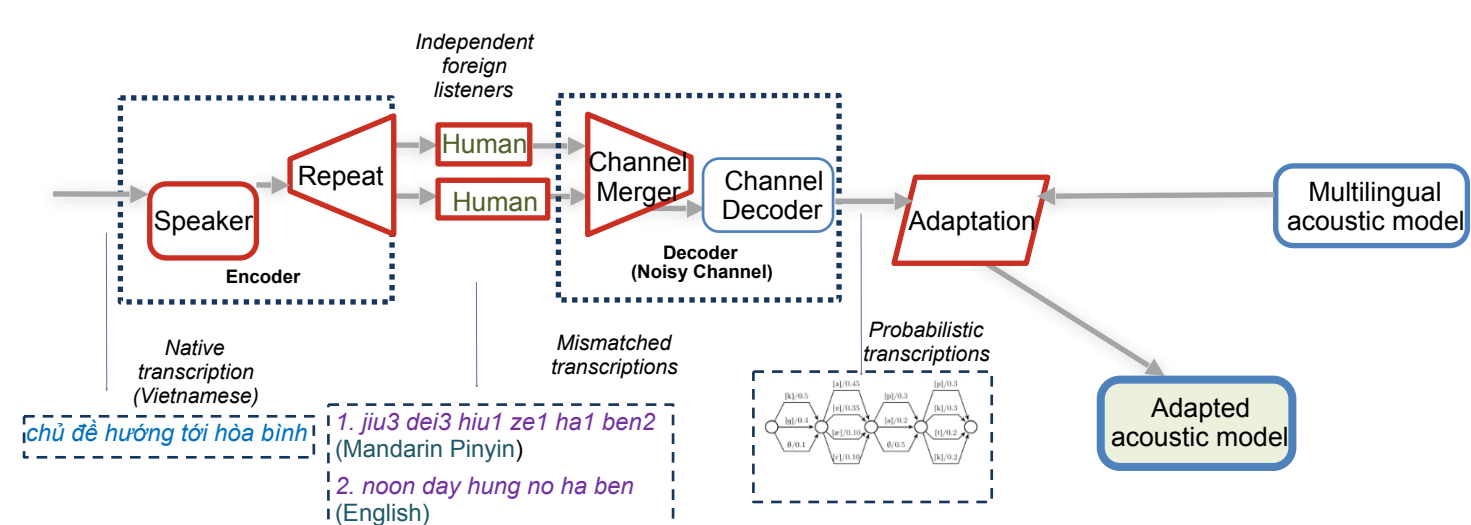
RECOGNIZING ZERO-RESOURCED LANGUAGES BASED ON MISMATCHED MACHINE TRANSCRIPTIONS

Wenda Chen^{1,2}, Mark Hasegawa-Johnson¹, Nancy F. Chen²

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

Overview

- Mismatched crowdsourcing based human probabilistic transcription (PT) has been proposed recently for training and adapting acoustic models targeting zero-resourced languages where we do not have any native transcriptions.
- 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 describe 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.
- We also present a fully automated unsupervised approach for zero-resourced speech recognition using mismatched machine transcriptions for transfer learning of phone models.



Introduction and Objectives

- Sample utterance: Vietnamese (Original):
Cũng cho biết chẳng hạn như là
- Label 1: Vietnamese in English Recognizer:
EH K IH N CH AA UW B EH IH K AA CH AA N H EH IH N N UW L AH
- Label 2: Vietnamese in Mandarin Recognizer:
GEN1 ZHUO1 MI1 GE1 HAI3 YA2 YOU1 LEN1
- Label 3: Predicted Transcripts (Vietnamese segment clusters):
k u N t s O b i t s a N h a : n J 1 l a :

$$w_{ij} = \frac{1}{N} \sum_{q \in \mathcal{X}_i} S_q(j) \Rightarrow \sigma_2 \{ D_X^{-\frac{1}{2}} W D_Y^{-\frac{1}{2}} \}$$

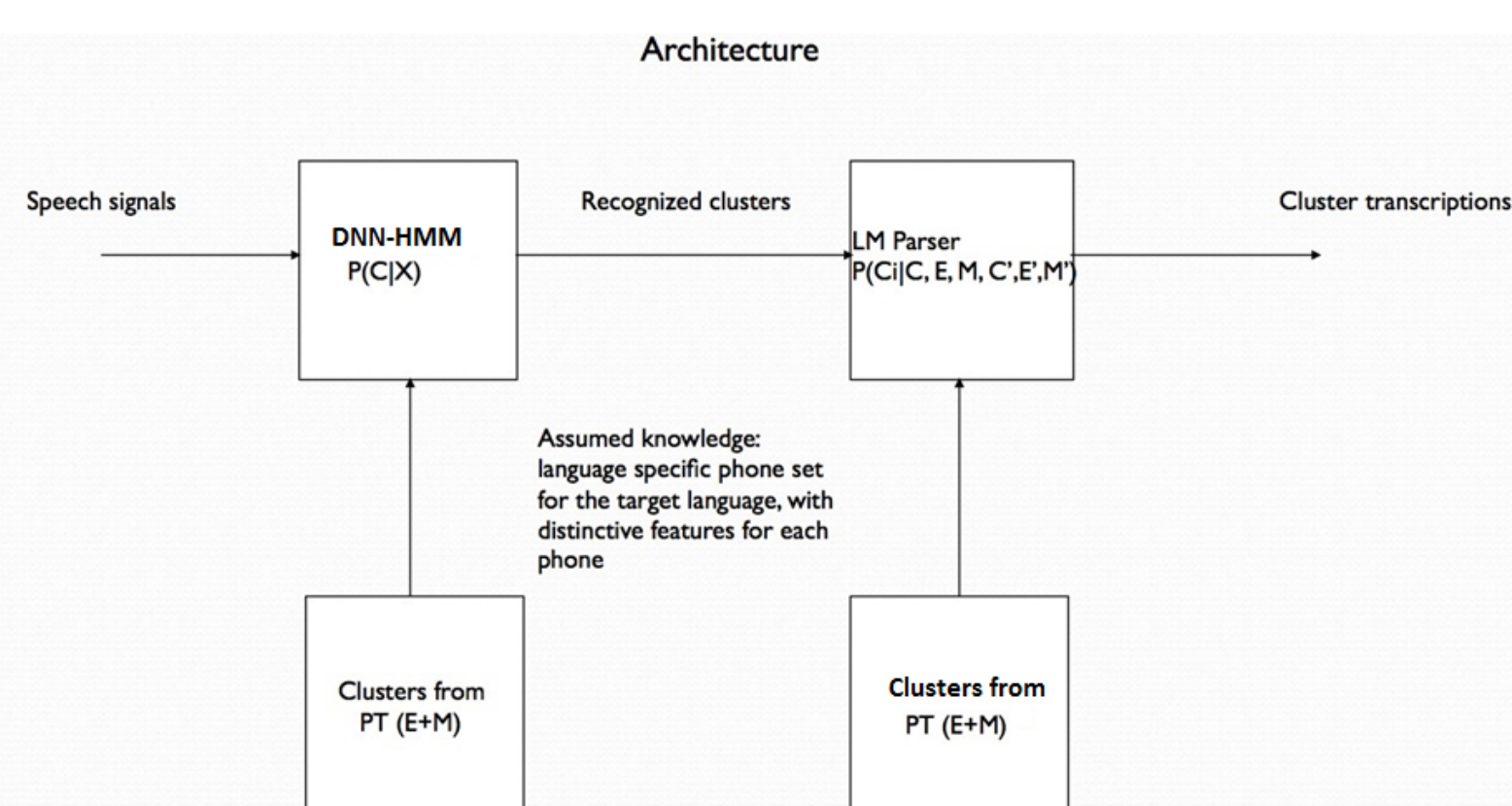
where $S_q(j)$ is the substitution probability for Mandarin phoneme j by English transcription token q , \mathcal{X}_i is the set of all transcription instances of the i^{th} English grapheme, and N is the number of all transcription segments in the training data.
 D_X and D_Y are the diagonal matrices where each diagonal element is the sum of the corresponding row or column of W .

Machine Mismatched Transcription Algorithm

- Step 1. Recognize** the target speech using English and Mandarin phone recognizers, such as the BUT phone recognizers and I2R speech recognizers, respectively. We collect the word level outputs from multiple available word recognizers, such as Google, CMU Sphinx, BUT, I2R, as the different machine transcribers and then convert them to phone level sequences using lexicons. The results will be compared with languages to find the more generalizable one to better recognize the target language. In this work, 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 and then derive the clusters using the clustering process. This makes use of the distinctive feature knowledge to characterize the phone differences between the languages. 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.

Modular Phone Recognition System

Proposed Modular System for Phone Recognition and Language Modeling (PRLM)



Machine Trans vs. Human Trans

Test languages are Vietnamese and Singapore Hokkien (Hokkien).

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%

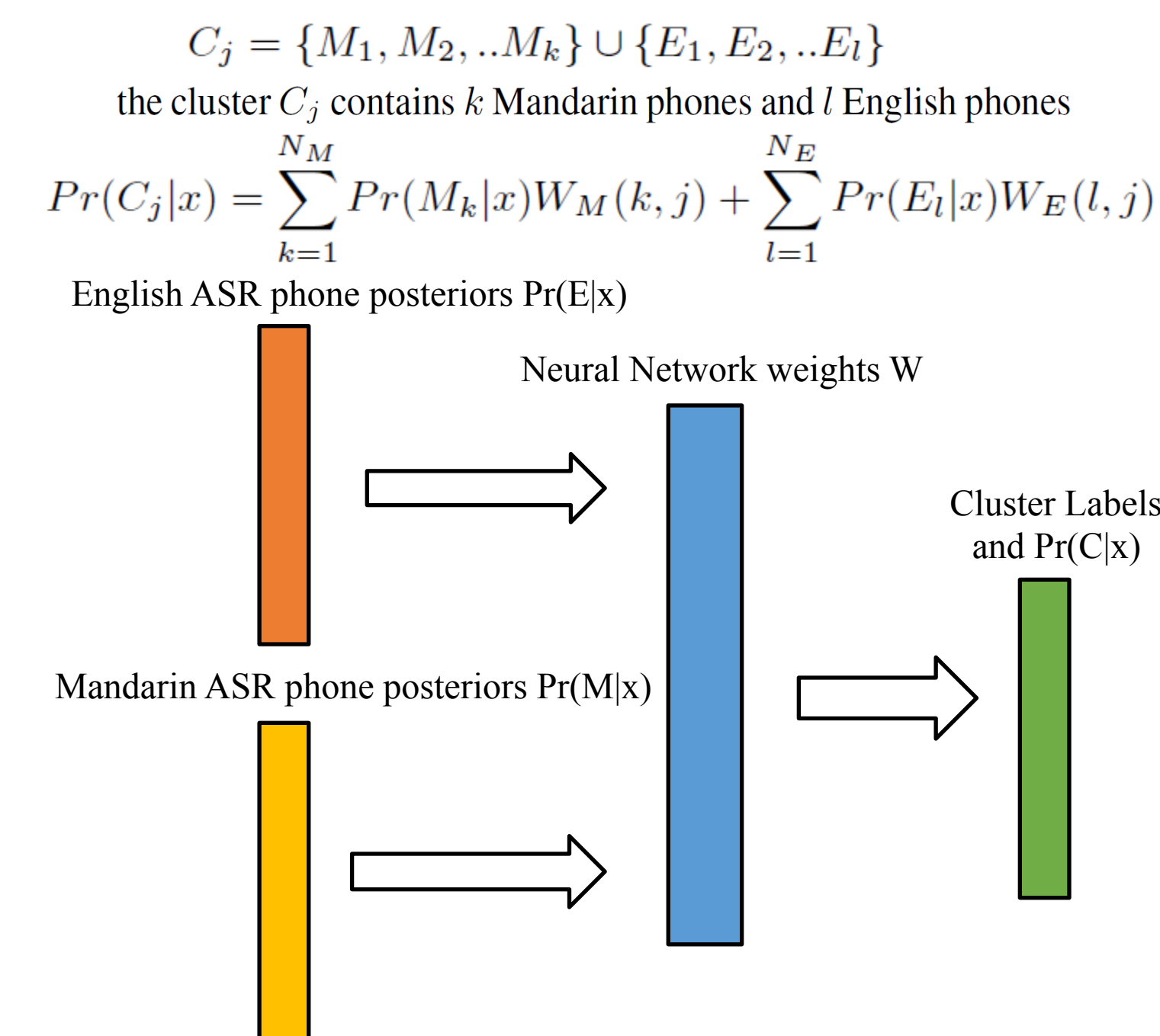
Table 2. Phone Error Rate (PER) of different recognition systems from BUT (Hungarian, English, Czech, Russian), I²R (English and Mandarin), and CMU Sphinx (Mandarin)

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. Clustering(Machine) is the clustering of the outputs of Hungarian and Mandarin recognizers.

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, and the machine system performance is relatively better than human based system performance on Hokkien but not on Vietnamese, we can guess that the threshold for deciding using human or machine systems could be **below 74% phone error rate**.

Formulations and Model



Proposed PRLM system: $p(C_t | \hat{C}_t, E_t, M_t, C_{t-1}, E_{t-1}, M_{t-1})$

Baseline 1 (Model based transfer learning): MAP adaptation system using PT
Baseline 2 (Semi-supervised self training of acoustic models): Cluster-trained Phone Recognition System

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.

The final system works better than the baselines for **both languages**.

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.
- 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.
- 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.
- Clustering method together with machine transcriptions can be formulated as an automated mismatched phone recognition system. The phone recognizer followed by phone language model is proposed. The exact machine vs. human threshold range and its generalizability need further studies on **more languages**.

Current and Future Work

