Utterance Classification in Speech-to-Speech Translation for Zero-Resource Languages in the Hospital Administration Domain

Lara J. Martin, Andrew Wilkinson, Sai Sumanth Miryala, Vivian Robison, and Alan W Black
Language Technologies Institute, Carnegie Mellon University

Introduction

Problem: Communities of people who *do not* speak English need to receive medical care at the University of Pittsburgh Medical Center (UPMC). Traditional translation services are expensive and difficult to find for less-common languages.

Goal: Build a speech-to-speech translation system to enable conversation between patients and hospital staff…
- For **any source language**, assuming **no previous data or linguistic knowledge**
- That requires only one bilingual speaker, for initial translation
- That is extensible to any other limited domain
- That runs in real time

Phonetic Representations of Audio

A. Detected English Phonemes [Sitaram et al.]:
Original Phrase: "What brings you here today?"
Eng. Phones: "SIL W AH T P R IH NG Z IY HH IH R D IH D EY"

B. Inferred Phonemes (IPs) [Muthukumar and Black]:
Original Phrase: "What brings you here today?"
47 IPs: "ip25 ip26 ip26 ip26 ip25 ip4 ip13 ip13 ip13 ip13 ip14 ip14 ip14 ip14 ip24…"

Building the System

1. Define template phrases
102 English phrases extracted from UPMC staff interviews
Examples: "Are you her legal guardian?"
"What brings you here today?"
"I'd like to reschedule my appointment."
"Do you have insurance?"

2. Acquire initial translations
One bilingual speaker translates the English phrases into the source language and records themselves speaking all the phrases

3. Acquire training data
Other source-language speakers record themselves repeating each recorded phrase from the original speaker
- 12 native English speakers (7 male, 5 female)
- 5 native Tamil speakers (5 male)

4. Learn to match new source-language utterances to template phrases

5. Classify
At runtime, repeat until dialogue ends:
- Source-language speaker talks; utterance is classified as matching one of the template phrases
  → corresponding target (English) phrase is played
- English speaker responds; utterance is classified to its closest template phrase
  → corresponding source-language translation is played

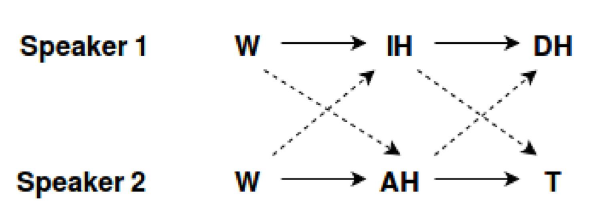
Ways of Classifying Phrases

Match an utterance from a new speaker to its most likely phrase, given a corpus of training data consisting of utterances over *n* template phrases spoken by *m* speakers

1. MFCC Dynamic Time Warping
   - Language-independent, acoustically derived
   - Slow, computationally expensive
   - Low accuracy

% Top 1	% Top 5	% Top 10	Avg. Rank
7.108	16.585	21.814	57.453

2. Logistic Regression
   - Binary features of cross-speaker bigrams of English phonemes



Language	% Top 1 Stage I	% Top 1 Stage II	% Top 10	Avg. Rank*
English	71.86	69.75	99.93	1.62
Tamil	34.51	41.6	71.38	2.82

3. String Edit Distance—Gaussian Model
   - Utterances of the same phrase across different speakers have their Levenshtein SED scores averaged, to create a model
   - In testing, each phrase is compared to each model by its z-score, and all scores are sorted

English

Feats	% Top 1	% Top 5	% Top 10	Avg. Rank
Eng	**87.173**	98.856	99.265	1.192
IP-15	80.310	99.265	99.265	1.238
IP-17	76.389	99.265	99.265	1.307
IP-30	80.474	99.183	99.265	1.262
IP-47	77.451	98.693	99.265	1.314
IP-84	79.248	98.856	99.265	1.316
IP-92	78.023	99.020	99.265	1.314
IP-101	79.575	98.611	99.265	1.312
IP-123	75.654	98.693	99.183	1.356
IP-171	76.716	98.693	99.265	1.347

Tamil

Feats	% Top 1	% Top 5	% Top 10	Avg. Rank
Eng	59.216	92.941	99.020	2.409
IP-14	60.392	98.627	99.608	1.821
IP-17	57.451	97.843	99.608	1.925
IP-24	59.608	97.647	99.804	1.749
IP-51	62.745	97.255	99.608	1.846
IP-82	62.157	97.255	99.608	1.755
IP-93	60.980	97.255	99.608	1.795
IP-103	58.824	98.039	99.608	1.839
IP-118	60.392	97.255	99.804	1.745
IP-165	**62.941**	95.882	99.412	1.820

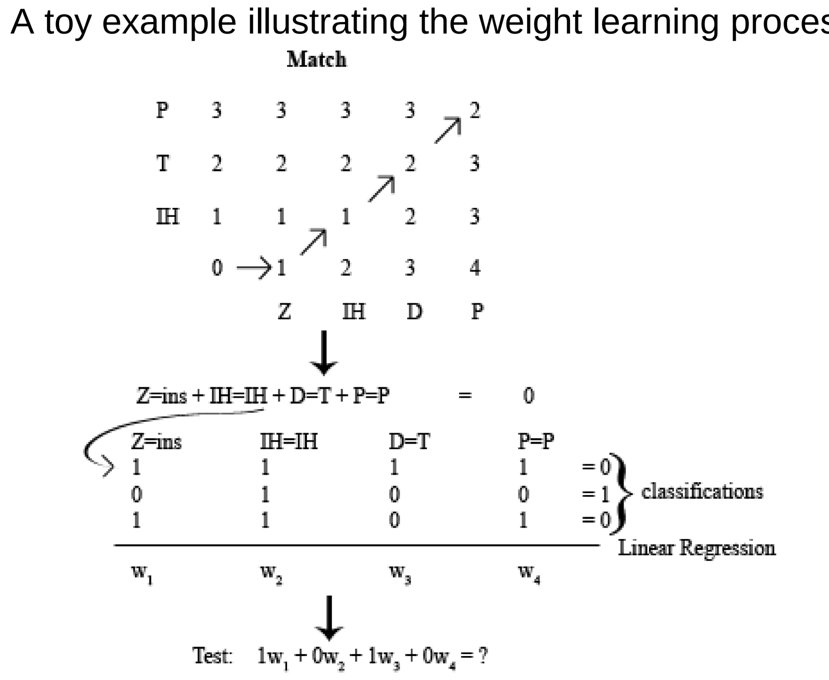
4. String Edit Distance—Phonetic Weights
   - Linguistic properties used to create table of weights for phoneme pairs

Language	% Top 1	% Top 5	% Top 10	Avg. Rank
English	59.722	97.386	99.020	1.724
Tamil	39.020	88.431	97.843	2.845
Eng-Small	21.242	67.974	78.758	14.620

5. String Edit Distance—Articulatory Feature Weights
   - Uses vector representations of AFs derived during IP discovery
   - Euclidean distance

Language	% Top 1	% Top 5	% Top 10	Avg. Rank
English	63.562	99.183	99.265	1.525
Tamil	37.843	94.510	99.216	2.714
Eng-Small	26.471	72.876	82.680	12.368

A toy example illustrating the weight learning process



6. Learning Weights—Iterations

English-Small

Iterations	% Top 1	% Top 5	% Top 10	Avg. Rank
No Weights	80.392	87.255	88.235	7.072
1x	**88.235**	**95.425**	**96.732**	**1.913**
2x	79.739	90.196	92.484	2.837
3x	76.797	87.255	92.157	3.433
4x	79.412	90.850	93.137	3.047
5x	76.797	89.216	92.484	3.700
6x	81.046	91.176	93.464	3.187

7. Learned Weights—One Iteration

English-Small

Feats	% Top 1	% Top 5	% Top 10	Avg. Rank
Eng	**88.235**	**95.425**	**96.732**	**1.913**
IP-15	13.072	34.641	48.039	27.437
IP-17	15.033	40.523	55.556	22.447
IP-30	21.895	49.020	60.458	16.623
IP-47	16.667	47.712	58.824	19.823
IP-84	18.627	42.157	55.556	20.077
IP-92	19.608	41.503	56.863	18.210
IP-101	15.033	37.908	49.673	20.467

Tamil

Feats	% Top 1	% Top 5	% Top 10	Avg. Rank
Eng	49.608	77.255	86.078	8.128
IP-14	25.098	53.529	65.882	13.736
IP-17	33.529	65.098	76.471	10.532
IP-24	37.843	70.784	83.137	6.375
IP-51	47.647	75.686	86.275	5.816
IP-82	47.647	74.510	82.745	7.000
IP-93	46.275	76.471	84.510	6.569
IP-103	49.608	76.667	84.314	6.413

Discussion

- English phonemes work the best with English
  - SED weight learning through linear regression
- For Tamil, the IPs work somewhat better than the English phonemes & learning SED weights does not improve results using the default weights
- To save time, we should use the English phones
- Attempt combining methods, in addition to other classification techniques

Questions:
- How many speakers are needed to build an adequate system, and which ones are most useful?
- Will our system be able to extend to our 750-plus phrases? Our 102 phrases are fairly phonetically distinct.
- Will dialogue-state tracking improve performance?

For references, please see paper.