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

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

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

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

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

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

6. Learning Weights—Iterations

7. Learned Weights—One Iteration

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.