Code-Switching Speech-To-Text

Created	@May 23, 2022 10:13 AM
Tags	

Code-switching ASR systems generally use the following:

- Acoustic model: hidden markov model (HMM)
  - or connectionist temporal classification (CTC)
- Language model (e.g. trigram) for phonetic-to-text (PTT)
- Sometimes a language boundary detection (LBD) system or a language identifier (LID)

This company appears to operate an existing trilingual solution. (Not clear whether this includes code switching or just supporting three languages separately).

Chinese-English Mixlingual Automatic Speech Recognition System (Hua et al.)

https://s3-us-west-2.amazonaws.com/secure.notion-static.com/578e0cf6-fc34-4317-bb6f-484194a7dd67/Chinese-English_Mixlingual_ASR.pdf

- Used SEAME dataset (code-switching audio of Mandarin and English with transcripts) as well as dictionaries for English and Chinese
- For acoustic features: extract MFCC or MFSC features, or Linear Discriminant Analysis and Maximum Likelihood Linear Transform (LDA+MLLT), or Speaker Adaptive Training (SAT)
- DNN used to predict the probability of a phone class given a specific feature (dependent on the Gaussian Mixture Model (GMM) -HMM trained on LDA feature to force alignment to get targets for each training frame)
- CNN achieves lowest MER but SGMM (Subspace Gaussian Mixture Model) just as good but much quicker to train
- Used trigram language model for PTT

Features

- MFCC
  1. Pre-emphasis filter
  2. Hamming window
  3. Fast Fourier Transform
  4. Convert to Mel scale
  5. Discrete Cosine Transform (DCT)
     a. Linear so not great for non-linear samples
- MFSC
  - Same as MFCC but without DCT
  - Better for NN approaches
- Incorporate first and second differential parameters to weaken HMM conditional independence assumption

- Linear discriminant analysis (LDA) + Maximum Likelihood Linear Transform (MLLT)
  - Supervised generative model to maximise ratio of class variance to average within class variance
  - Reduces number of input dimensions
- Speaker Adaptive Training (SAT)
  - Projects onto speaker-neutral space
  - Reduce wasted time learning about speaker style as opposed to content

OC16-CE80: A Chinese-English Mixlingual Database and A Speech Recognition Baseline

https://s3-us-west-2.amazonaws.com/secure.notion-static.com/92c0c6eb-293a-45ad-a353-60ccc69698e2/1609.08412.pdf

- Uses similar Mandarin-English code-switching dataset as well as dictionaries
- GMM-HMM using MFCCs produced forced alignment for training data
- Training data used in DNN-HMM (TDNN)
- Language model with trigrams (best when trained on mixlingual dataset as well as a monolingual dataset)

Towards Code Switching ASR for E2E CTC Models

https://s3-us-west-2.amazonaws.com/secure.notion-static.com/d2195461-1525-4e09-ac53-1d3311ce5b1c/Towards_code_switched_ASR_for_End_to_End_CTC_models.pdf

- Use bi directional LSTM in CTC to output the token
- Separate frame-level language ID (LID) model to identify the language over a short period
- Use the output from the LID to select the correct output from the CTC
- Best performance gained when CTC initialised from being trained on primary language, and then further trained on context-switching examples

Automatic Recognition of Cantonese CS Speech

https://s3-us-west-2.amazonaws.com/secure.notion-static.com/28520840-ffd1-44b0-a522-7114464c3925/O09-5003.pdf

- Only Cantonese-English example I could find
- Note that pronunciation is different in context-switching speech than in monolingual speech (syllables can be pronounced differently due to different syllable/vowels/consonants in the two languages. Hence using just two monolingual models is not effective
- Acoustic model using HMM from features
- Mix monolingual speech databases TIMIT and CUSENT with CUMIX (context-switching database) for best training

- Train trigram language model phonetic-to-text (PTT)
  - Used articles containing specific characters considered indicative of spoken Cantonese as opposed to standard Chinese
- Language boundary detection (LBD) based on the syllable graph output by the acoustic model to rescore before passing the syllables through the PTT
  - Different length of syllables in English and Cantonese
  - Syllable graph edge probabilities edited based on the hypothesised language

Links

TIMIT Acoustic-Phonetic Continuous Speech Corpus

The TIMIT corpus of read speech is designed to provide speech data for acoustic-phonetic studies and for the development and evaluation of automatic speech recognition systems. TIMIT contains broadband recordings of 630 speakers of eight major dialects of American English, each reading ten phonetically rich sentences.

https://catalog.ldc.upenn.edu/LDC93s1

SHACHI - Language Resource Metadata Database

CUSENT is a part of CUCopora, a large scale Cantonese spoken language corpora. It is a large collection of spoken Cantonese sentences designed to be phonetically rich. The corpus also includes manually verified phonemic transcription.

http://shachi.org/resources/3269

Other databases | Technology Licensing | DSP Lab

CU2C and CUMIX were developed at the DSP and Speech Technology Laboratory, Department of Electronic Engineering, CUHK. CU2C is a dual-condition Cantonese speech database for speaker recognition research. It is a task-oriented database. The speech contents include Hong

http://dsp.ee.cuhk.edu.hk/license_otherdatabases.php

CMU Pronouncing Dictionary

The pronunciation of over 134,000 North American English words

https://www.kaggle.com/datasets/rtatman/cmu-pronouncing-dictionary?select=cmudict.vp

Speech Recognition (ASR) and Speech-to-Text (STT) Cantonese Chinese Company in Hong Kong

Best and highest-quality廣東話,中文Cantonese Speech Recognition (ASR), Speech-to-Text (STT), and Voice Recognition provider/manufacturer for companies and businesses in Hong Kong. Multilingual ASR in Cantonese, Putonghua Chinese, English (廣東話, 港式粵語, 普通話, 中文, 英語) 語音控制 | 語音辨識 | 文語轉換

https://www.infotalkcorp.com/speech-recognition-asr-stt/

Conclusions

General consensus is to follow the following steps:

1. Extract features (MFCC, MFSC, LDA+MLLT, SAT)
2. Estimate probabilities of phones
3. Use language model to predict text from the pronunciation (phones) probabilities

Somewhere in there, some approaches use a language identifier or language boundary detector to work out which frames belong to which language and rescore them (weight the probabilities of the different phones) appropriately to improve output.

Existing solutions generally support multiple languages separately with different models (including AWS). A very simple solution could be to use a language boundary detector to work out where one language begins and the other ends, and then pass these snippets into the correct language model. This may work for some speakers, but if the accent and style of speech is carried over between languages, it is unlikely to work beyond simple, clear audio samples. This is because the pronunciation/phones are morphed towards the primary language. Additionally boundary detection may not be accurate enough, because a lot of the time the language boundary is determined by the meaning more than the sound.