Integration of unsupervised acoustic, lexicon, and language models toward language acquisition from speech

Shinji Watanabe

JHU mini workshop
July 2012
My research background

• Automatic speech recognition
  – Acoustic modeling (mainly): Bayesian acoustic modeling, adaptation, discriminative training
  – Language modeling (sometimes): topic tracking language model
  – Discriminative model for WFST based ASR decoder

• I recently started unsupervised (zero resource) spoken language processing
  – Good application of Bayesian approaches
  – It’s ongoing research
Hierarchical dynamics in speech (recognition)

Frame unit (10ms) \( o_1, o_2, o_3, o_4, o_5 \)
HMM state unit (~50ms)
Phoneme unit (~0.1s) /w/, /a/, /t/, /a/, /n/, /a/, /b/, /e/
Word unit (~0.3s) Hello, my name is Watanabe.
N-gram unit (~1s) my+name+is, name+is+Watanabe,
Utterance unit (~5s) Hello
Or chunk unit (~1m) My name is Watanabe

Signal processing
Feature extraction
Acoustic model
Lexical model
Language model

There are topic transition
speaking style changes, speaker changes
Strategy

• Represent a dynamics on each layer of speech with a straightforward generative model
  – Phoneme: HMM, word: n-gram
  Basically using automatic speech recognition techniques
5+ layer speech generative model

Speaker change dynamics
Topic dynamics

Word n-gram
Phoneme n-gram
Left to right
HMM
GMM
Speech feature (MFCC)
WFST based representation

- Hidden Markov Model
  - MFCC seq. → phone
  - Acoustic model

- (Probabilistic) Regular grammar
  - phone → word
  - Lexicon model

- (Multiple) Markov model
  - word → sentence
  - Language model

WFST based representation
WFST based representation

Acoustic model

Lexicon model

Language model

Optimized WFST

Hidden Markov Model

(Multiple) Markov model

(Probabilistic) Regular grammar

Composition

MFCC seq. → phone

phone → word

word → sentence

MFCC seq. → sentence
WFST based representation

- Flat representation
  ⇒ makes inference algorithm simple

- Optimized WFST

- MFCC seq. → sentence
Inference algorithm (skip details)

- Acoustic modeling:
  - Segmental k-means (ML-EM) or utterance-unit blocked Gibbs (not a nonparametric Bayes) using WFST

\[
p(O, Z, S, V, W | \Theta) = p(w_{\text{end}} | w_N) \prod_{n=1}^{N} p(w_n | w_{n-1}, w_{n-2}) p(z_{\text{end}}^{n} | z_{L+n+1-1})
\]

\[
\prod_{l=L_n}^{L_n+1-1} p(z_l | z_{l-1}, w_n) p(s_l^{\text{end}} | s_{T_l+1-1}) \prod_{t=T_l}^{T_{l+1}-1} p(s_t | s_{t-1}, z_l) p(v_t | s_t) p(o_t | v_t)
\]

where

\[
p(w_n | w_{n-1}, w_{n-2}) = \begin{cases} p(w_n) & n = 1 \\ p(w_n | w_{n-1}) & n = 2 \\ p(w_n | w_{n-1}, w_{n-2}) & n > 2 \end{cases}
\]

\[
p(z_l | z_{l-1}, w_n) = \begin{cases} p(z_l | w_n) & l = L_n \\ p(z_l | z_{l-1}) & L_n < l < L_n+1 \end{cases}
\]

\[
p(s_t | s_{t-1}, z_l) = \begin{cases} p(s_t | z_l) & t = T_l \\ p(s_t | s_{t-1}, z_l) & T_l < t < T_{l+1} \end{cases}
\]
Inference algorithm (skip details)

• Acoustic modeling:
  – Segmental k-means (ML-EM) or utterance-unit blocked Gibbs (not a nonparametric Bayes)

```latex
Algorithm 2 Unsupervised language acquisition.

1: Initialize $\Psi \leftarrow \Psi^0$
2: Sample $Z \sim p(Z|\Psi)$
3: for $\tau = \{1, \ldots \}$ do
4:   Compute $\{\Omega^u\}_{u=1}^U$
5:   for $u = \text{shuffle}(1, \ldots, U)$ do
6:     Update $\Omega^u$
7:     Update $\Psi \leftarrow \Psi^u$
8:   Prune $q(Z^u)$ from $p(Z^u|\Psi)$
9:   Sample $Z^u \sim q(Z^u)$
10:  Compute $\Omega^u$ from $Z^u$
11: end for
12: Sample $W \sim p(W|O, Z)$
13: Sample $Z \sim p(Z|O, W)$
14: end for
```

• Lexicon and language modeling:
  – Gibbs sampling based on Hierarchical Pitoman-Yor Process [Mochihashi (2009), Neubig (2010)]
LatticIm (Graham Neubig)

• Open source tool for word segmentation based on HPY
  – Input: phoneme sequences for “phoneme lattices”
  – Output: word sequences
• Implemented based on openfst
  ⇒ easily integrated with other components
Flow chart

- MFCC
- Randomly initialized pseud phoneme labels
- AM
- Clustered phoneme labels (lattices)
- WFST
- LM (latticelm)
- Word (?) segmentation results
Examples of pseud phoneme labels

- Random initialization
  - Uniformly samples numbers from $[1p, 2p, \ldots, 48p]$. The length of phoneme sequence is proportional to the length of utterance
  - Beginning and end phonemes are fixed as “$1p$” <-silence

- After unsupervised AM
  - Segmental k-means

- After unsupervised LM
  - Concatenating phonemes to obtain word sequences
Flow chart (feed back from LM)

MFCC

Randomly initialized pseud phoneme labels

AM

Clustered phoneme labels (openfst lattices)

WFST

LM (latticelm)

Word (?) segmentation results
Openfst format of lattices

- Weight: acoustic score (minus log likelihood) + language scores (minus log probability with a scaling factor)
Experiments

“Preliminary experiments”
Experimental condition

- TIMIT female training set (112 utterances, 1088 utterances)
- Acoustic modeling (Kaldi: tightly integrated with openfst)
  - Acoustic condition
    
    | Sampling Rate | 16 kHz |
    | Quantization  | 16 bit  |
    | Feature Vector| 12 - order MFCC with log energy with Δ and Δ Δ |
    | Window        | Hamming |
    | Frame Size/Shift | 25/10 ms |

  - Acoustic model
    - **48, 100, 500, 1000** phonemes
    - 3 state left-to-right HMMs
    - 8 Gaussian mixture components
    - Phoneme bigram:
    - Lexicon and language modeling (latticelm)
      - Phoneme 3gram, Word 3gram
Performance criterion

- Diarization error rate
  - Who speaks when?

\[ \text{DER} = \frac{\sum_{s=1}^{S} \text{dur}(s) \cdot (\max(N_{\text{ref}}(s), N_{\text{hyp}}(s)) - N_{\text{correct}}(s))}{\sum_{s=1}^{S} \text{dur}(s) \cdot N_{\text{ref}}} \]

- Changing from speaker clusters to phoneme clusters
  - Phoneme alignment (obtained by Viterbi algorithm)
Experimental results (unsupervised AM)

- It seems to work to some extent, but not good performance
- Gibbs would be better (mitigating the local optimum problem (?)), but it takes so much time
Example (* different experimental setup)

I have not tried any objective evaluation for word segmentation yet.

Some basic syllables seemed to be extracted (?)

So many errors were included in this conversion
  - Errors in unsupervised acoustic and language modeling
  - Token to phoneme mapping

Feed back from ULM to UAM currently did not improve performance
Summary

- Unfortunately the approach did not work well due to bugs? mistakes?, but it would be a good starting point for us toward language acquisition from speech
- I don’t use any label information at all, but use some knowledge
  - Average phoneme length, number of phonemes, ASR knowledge
- Unsupervised acoustic modeling should be improved
  - Feature, model, training method
- Feed back part is not tightly integrated (not fst)
- Evaluation measure
  - DER would not be a good measure
- Model complexity control in acoustic modeling
Tools

• Kaldi (ASR tool)
  http://kaldi.sourceforge.net/index.html

• Openfst
  http://www.openfst.org/twiki/bin/view/FST/WebHome

• Latticelm
  https://github.com/neubig/latticelm

• Diarization error rate
My final goal

Unsupervised generative models of whole speech communication

- Who speak when, what?
- What is a topic?
- What kind of room environment, atmosphere, role, emotion

Learning generative models involving everything related to speech communication
Thank you for your attention