Speech Recognition with Segmental Conditional Random Fields

Geoffrey Zweig, Patrick Nguyen, Dirk Van Compernolle, Kris Demuynck, Les Atlas, Fei Sha, Hynek Hermansky, Aren Jansen, Damianos Karakos
The Team!

Senior Members
- Les Atlas, University of Washington
- Kris Demuynck, Leuven University
- Hynek Hermansky, JHU
- Aren Jansen, JHU COE
- Damianos Karakos, JHU
- Patrick Nguyen, Microsoft Research
- Fei Sha, USC
- Dirk Van Compernolle, Leuven
- Geoffrey Zweig, Microsoft Research

Student Members
- Sam Bowman, University of Chicago
- Pascal Clark, UW
- Sriram Ganapathy, JHU
- Sivaram GSVS
- Justine Kao, Stanford
- Greg Sell, Stanford
- Samuel Thomas, JHU
- Meihong Wang, USC

Thanks!
- Brian Kingsbury
- Ken Church

6/21/2010
Goal

- Advance the state-of-the-art in core speech recognition
  - Use segmental CRFs to gain the advantages of long-span log-linear models in a way appropriate for continuous speech recognition

Observations blocked into groups corresponding to words. Observations typically detection events.

States represent whole words (not phonemes)

Log-linear model relates words to observations

$O_1$  

$O_n$
Why Segmental CRFs?

- SCRFs enable us to unify the application of powerful new scientific approaches to ASR—e.g.
  - Template detections [Van Compernolle et al. 03]
  - Auto-encoding neural net features [Mohammed & Hinton 09]
  - Coherent modulation features [Atlas 09]
  - Point Process word models [Jansen 10]
  - Sparse Representation Phoneme Detectors [Hermansky et al. 10]
- No loose ends
  - Language model built-in, joint discriminative AM/LM training
  - Training, decoding processes suitable for LVCSR
  - Source code available & extensible by community

6/21/2010
SCARF Introduction
- Patrick Nguyen -
Model Structure

States represent whole words (not phonemes)

Log-linear model relates words to observations

Observations blocked into groups corresponding to words. Observations typically detection events.

For a hypothesized word sequence \(s\), we must sum over all possible segmentations \(q\) of observations

\[
P(s|o) = \frac{\sum_q \text{s.t. } |q|=|s| \exp\left(\sum_{e \in q} \lambda_k f_k(s_{le}^e, s_{re}^e, o(e))\right)}{\sum_{s'} \sum_q \text{s.t. } |q|=|s'| \exp\left(\sum_{e \in q} \lambda_k f_k(s_{le}^{e'}, s_{re}^{e'}, o(e))\right)}
\]

Training done to maximize product of label probabilities in the training data (CML).

6/21/2010
The Observations

- Detector streams
  - Phone detectors
  - Syllable detectors
  - Multiphone detectors
  - Template detectors
- Could also use raw speech frames
- At the end of the day, just need to define features that measure the consistency between a word hypothesis and the underlying acoustics
Inputs (1)

- Atomic streams
  - (detection time) +
- Optional dictionaries
  - Specify the expected sequence of detections for a word
Inputs (2)

- Lattices to constrain search
The Features

- Array of features automatically constructed
- Measure forms of consistency between expected and observed detections
  - Differ in use of ordering information and generalization to unseen words
- Existence Features
- Expectation Features
- Levenshtein Features
- Also
  - LM features
  - “Baseline” feature
  - User-defined features

6/21/2010
Existence Features

- Does unit X exist within the span of word Y?
- Created for all X,Y pairs in the dictionary and in the training data
- No generalization, but arbitrary detections OK

Hypothesized word, e.g. “accord”

Spanned units, e.g. “ih k or”
Levenshtein Features

- Match of u
- Substitution of u
- Insertion of u
- Deletion of u

Align the detector sequence in a hypothesized word’s span with the dictionary sequence that’s expected

- Count the number of each type of edits
- Operates only on the atomic units
- Generalization ability across words!

6/21/2010
User-Defined Features

```
<10.16.100C2934-9090-453E-890B-2D96FC60D7BC.dc
1 90 <s> myfeat=-0.035779,offset=1.0
91 1090 [dtmf] myfeat=0.000000,offset=1.0
91 1090 [fragment] myfeat=-0.059355,offset=1.0
91 1090 zoned myfeat=-2.694036,offset=1.0
91 1250 bleu myfeat=-1.370601,offset=1.0
91 1250 block myfeat=-1.485341,offset=1.0
91 1250 blu myfeat=-1.329818,offset=1.0
91 1250 blue myfeat=-1.328225,offset=1.0
91 1330 bleu myfeat=-0.802841,offset=1.0
91 1330 bloom myfeat=-0.828402,offset=1.0
91 1330 blu myfeat=-0.810672,offset=1.0
91 1330 blue myfeat=-0.789534,offset=1.0
91 1330 blues myfeat=-0.835714,offset=1.0
91 1446 bloom myfeat=-0.921589,offset=1.0
91 1446 blue myfeat=-1.100250,offset=1.0
91 1446 blues myfeat=-0.941625,offset=1.0
91 1446 lube myfeat=-1.249081,offset=1.0
91 1446 lufe myfeat=-0.938143,offset=1.0
91 1562 bluebird myfeat=-0.575483,offset=1.0
```
In Workshop We’ll Make:

Symbolic Streams

Lattice Annotations And Cohort Lattices
Information Sources Overview
- Geoffrey Zweig Summarizing-
Research Themes

SCARF Glue

Integrates multiple information types – standard systems don’t

Neural Net

Modulation Features

Template

Training

6/21/2010
NN/STRF Features in SCARF

Activation Threshold

Detections

Symbolic Stream

$s_l$, $e$, $s_r$, $o(e)$

$O_1$, $O_n$
Deep learning for NNs

- Multilayer neural networks
  - Trained both generatively and discriminatively, ie, pre-training and fine tune
  - Trained greedily, ie, layer by layer
  - Each layer learns new correlations between the features
  - Topmost layer predicts class labels
NN Related Work at JHU

- Phonemes as indicators of acoustic events in SCARF
  - with Samuel Thomas and Keith Kintzley
  - HMM-NN hybrid
  - matched filters

- TANDEM-like and sparse features for baseline LVCSR
  - with Samuel Thomas and Sivaram Garimella
  - hierarchical phoneme estimates
  - autoassociative NN for deriving sparse features

6/21/2010
STRF Syllable Detector

- Inspired by Hynek Hermansky & Nima Mesgarani

In-Class Examples

Out-of-Class Examples

Max-Ent Classifier

STRF for Syllable: ME Weight Matrix

x 5000: One for each syllable
Point Process Models

Atomic detections

Poisson density model
E.g. “greasy”

Lattice annotation with Model scores
Template Based System(s) – Motivation & Concepts

• Motivation:
  • Do away with the 1st order Markov assumption
  • Work with longest possible units in a dynamic way (no predefined 3ph, 5ph, ..)
  • Exploit meta-information available in the original data, but discarded in the HMM model building phase
Template Detections in SCARF Database

Input

Matches

Symbolic Stream

6/21/2010
Semi-Supervised Discriminative Learning with SCARF

View ASR problem as Word Sense Disambiguation

1) Use untranscribed audio to learn frequent word confusions (cohort sets)
2) Use data with ground truth to learn to disambiguate these confusions
3) Modify lattices to reflect the cohort information
   a) Remove spurious non-cohort links
   b) Add missing links
4) Create specialized word detectors for cohort words

- Related work
  Xu, Karakos and Khudanpur (ASRU 2009)
  Contrastive estimation (Smith & Eisner 2005)
  Kurata et al. (ICASSP 2009)
  Huang, Li & Acero (ICASSP 2010)
Coherent Modulation

Claim: Speech signals encode information via low-frequency envelopes modulating high-frequency carriers

“Bird populations”
Coherent vs Incoherent Modulation

Assume non-negative envelope

Discontinuity in carrier:
→Carrier not bandlimited!

Assume bandlimited carrier

Temporal envelope will, in general, be complex
Coherent vs Incoherent Modulation

Synthesis - Coherent Modulation

Synthesis - Incoherent Modulation

6/21/2010
Coherent Modulation in SCARF

hypothesized word

Parametric representation of modulators

200 256 bird

\( m_k(n) \)

Model of what “bird” modulators should be

A Score

Lattice annotation with Model scores

6/21/2010
Data & Progress

- Very good Broadcast News baseline setup in place
  - Thanks Damianos for building it!
  - Thanks to Brian Kingsbury for help & advice
  - Thanks IBM for use of Attila
  - Improvements observed from using SCARF with Multiphone detectors, discriminative LM training and word detectors
    - We can move the needle!
- World-record level WSJ template system
  - Thanks Kris!
  - Now being SCARF’d

6/21/2010