Whole Word Discriminative Point Process Models

Aren Jansen

Human Language Technology Center of Excellence, Center for Language and Speech Processing
Johns Hopkins University, Baltimore, MD USA

ABSTRACT

This paper introduces a discriminative extension to whole-word point process modeling techniques. Meant to circumvent the strong independence assumptions of their generative predecessors, discriminative point process models (DPPM) are trained to distinguish the composite temporal patterns of phonetic events produced for a given word from those of its impostors. Using correct and incorrect word hypotheses extracted from large vocabulary recognizer lattices, we train whole-word DPPMs to provide an alternative set of acoustic model scores. Using solely the timing of sparse phonetic events, DPPM scores exhibit comparable discriminative power to those produced by a state-of-the-art acoustic model built using the IBM Attila Speech Recognition Toolkit. In addition, the inherent complementarity of frame-based and event-based models permits significant improvements from score combination.

MOTIVATION

A Surprising Fact

2/3 of the 430 hour HUB4+TDT4 broadcast news training corpus is covered by words with at least <u>1000</u> occurrences

An Opportunity:

Transition to acoustic models of entire words (demonstrated superior in <u>1989</u> by Lee et al.) for types with enough training examples

Potential Benefits:

1. Circumvent limitations of canonical pronunciation dictionaries for informal genres
2. Exploit longer range dependencies by avoiding the first-order Markov assumption

A Candidate: Point Process Models (PPM)

1. Transform the speech signal into sparse point patterns of salient acoustic/phonetic events in time
2. Explicitly model linguistic objects (e.g. words, syllables) with the temporal statistics of these patterns

Past Work: *Generative* PPMs [Jansen & Niyogi, 2009, 2010]

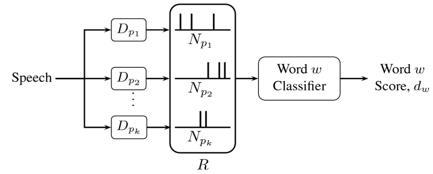
- **Core Model:** Predict word likelihoods assuming observed events are generated by an underlying word-dependent inhomogeneous Poisson process with time dependent intensity functions
- **Weakness:** Strong independence assumptions between events

Proposed Approach: *Discriminative* PPMs

1. Use LVCSR lattices to generate candidate word occurrences
2. Discriminatively train classifiers to distinguish correct and incorrect lattice arcs using entire temporal event patterns
3. Use classifier scores in LVCSR [Siu & Gish, 1999], term detection [Vergyri et al., 2006], or spoken dialog systems [Hazen et al., 2002]

DISCRIMINATIVE POINT PROCESS MODELS

High-Level Processing Architecture

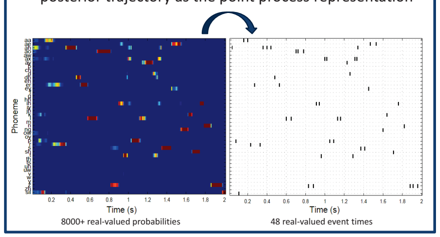


Two Primary Components:

1. A set of **phonetic event detectors** $\{D_p\}$ for each phone p of English
   - Each phone detector produces a set of event times $N_p = \{t_1, t_2, \ldots\}$ that phone p is most clearly produced
   - The collection of events set $R = \{N_p\}$ defines the point process representation
2. A set of **word detectors** $\{d_w\}$ for each word w that map a portion of R to a real-valued word confidence scores

MLP-Based Phone Detectors

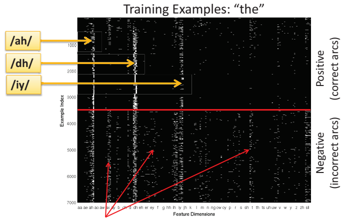
1. Compute multi-layer perceptron based monophone models using multi-stream methods [Thomas et al., 2009]
2. Threshold and take the **times of local maxima** of each posterior trajectory as the point process representation



Discriminative Word Models

1. Collect correct/incorrect lattice arcs for each word type and compute the phone event point pattern R for each
2. Transform each arc point pattern into a fixed dimensional vector of event counts in uniform time bins for each phone (+ duration)
3. Use your favorite machine learning algorithm to train a classifier for each word (we use kernel machines *a la* [Layton & Gales, 2007])

Training Examples: "the"



"Random" phone events present in negative examples only

EXPERIMENTS

Evaluation Corpus and Word Set

- 430-hour **Hub4+TDT4 broadcast news corpus**, split into two equal parts for training and dev/eval
- Evaluate **100 most confusable words** for the baseline recognizer (in order of descending errors):

  the and that to is in of are it on for had he you all with there as but what we was they them or have at about will not be up their out our when way this said now like an where think then some so one if how has good could your yeah why which were two time three more me his get do day because well than no into him here going down can am after yes would world who while war very too today thing see says right really over only off my most many long

- Examples drawn from Attila lattices and instances from the reference transcript forced alignment that Attila missed

Baseline Scores: IBM Attila Recognizer [Soltau et al., 2007]

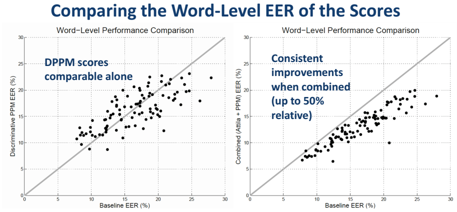
- **State-of-the-art acoustic model:**
  - ✓ Quinphone states, 150k Gaussian (total), LDA, VTLN, fMLLR, fMMI, MLLR, bMMI
- **Language model:** 400M word trigram model
- Lattices generated using both acoustic and language models
- Baseline word scores taken to be confusion network posteriors computed from the *acoustic model likelihoods only* [Mangu et al., 2000; Falvigna et al., 2002]

Discriminative PPM Scores:

- Train **regularized least squares** classifiers with **radial basis function kernel** for each confusable word (SVMs equivalent)
- Training examples per word range from 2k to 25k
- 421-dimensional feature vectors =

  [10 uniform time bins/phone] x [42 phones] + [1 duration]

Comparing the Word-Level EER of the Scores



Average EER Across the 100 Word Types

Score	Lattice Only	Lattice + Forced Alignments
Attila Confusion Network Posteriors	12.6%	17.1%
Generative PPM	16.1%	17.8%
Discriminative PPM	14.6%	16.4%
Attila + DPPM	**10.2%**	**13.4%**

1. **Discriminative PPM training adds extra power over generative PPM predecessor**
2. **DPPM scores can recover examples from the forced alignment missed by the Attila recognizer**
3. **Combining Attila confusion network posteriors with the DPPM score provides by far the best discrimination**

LVCSR INTEGRATION USING SCARF

JHU Summer Workshop 2010:

Integrate DPPM scores with Attila recognizer using Microsoft Research's Segmental Conditional Random Fields Toolkit (SCARF) [Zweig et al., 2011]

System	dev04f	rt04 (eval)
Baseline (Attila)	16.3% WER	15.7% WER
+ MSR HMM scores	15.3	14.5
+ DPPM scores (72 words)	15.0	14.3
(Lattice Oracle)	11.2	10.1

25% of possible reduction

System (all dev04f)	Unigram LM	Trigram LM
Attila + SCARF	16.9%	16.0%
Attila + SCARF + DPPM (72 words)	16.2%	15.8%

DPPM scores almost as beneficial as trigram LM

CONCLUSIONS

1. The community has amassed enough data in some languages to build whole word models for a surprisingly large number of word types
2. Whole word discriminative point process models provide a novel mechanism for producing alternative acoustic models scores
3. Discriminative PPM scores are competitive with state-of-the-art HMM-GMM acoustic models and offer a high degree of complementarity

ACKNOWLEDGEMENTS

The late **Partha Niyogi** contributed to this work and would have been a co-author if he were able to provide his consent. The author would like to thank **Damianos Karakos** of the Center for Language and Speech Processing (CLSP) at Johns Hopkins for his invaluable help in computing the baseline Attila lattices and confusion networks; **Samuel Thomas**, also of CLSP, for providing the MLP-based phonetic acoustic models; and **Geoffrey Zweig** of Microsoft Research for assistance in data preparation.

REFERENCES

D. Falavigna, R. Gretter, and G. Riccardi, "Acoustic and word lattice based algorithms for confidence scores," in *Proc. of ICSLP, 2002.*

T.J. Hazen et al., "Recognition confidence scoring and its use in speech understanding systems," *Comp. Speech and Lang.*, vol. 16, no. 1, pp. 49–67, 2002.

A. Jansen and P. Niyogi, "Point process models for spotting keywords in continuous speech," *IEEE Trans. Audio Speech Lang. Process., vol. 17, pp. 1457–1470, 2009.*

A. Jansen and P. Niyogi, "Detection-based speech recognition with sparse point process models," in *Proc. of ICASSP, 2010.*

M. Layton and M. Gales, "Acoustic modelling using continuous rational kernels," *J. VLSI Sig. Proc., vol. 48, 2007.*

C.-H. Lee et al., "Word recognition using whole word and subword models," in *Proc. of ICASSP, 1989.*

L. Mangu, E. Brill, and A. Stolcke, "Finding consensus in speech recognition: word error minimization and other applications of confusion networks," *Comp. Speech and Lang., vol.* 14, pp. 373–400, 2000.

M. Siu and H. Gish, "Evaluation of word confidence for speech recognition systems," *Comp. Speech and Lang., vol. 13, pp.* 299–319, 1999.

H. Soltau et al., "The IBM 2006 GALE Arabic ASR system," in *Proc. of ICASSP, 2007.*

S. Thomas, S. Ganapathy, and H. Hermansky, "Phoneme recognition using spectral envelope and modulation frequency features," in *Proc. of ICASSP, 2009.*

D. Vergyri et al., "The SRI/OGI 2006 spoken term detection system," in *Proc. of Interspeech, 2007.*

G. Zweig et al., "Speech recognition with segmental conditional random fields: A summary of the JHU 2010 summer workshop," in *Proc. of ICASSP, 2011.*