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 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 corpus training corpus is covered by words with at least 1000 occurrences

An Opportunity: Transition to acoustic models of entire words (demonstrated superior in 1989 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 phonetic/phonemic events in time
2. Explicitly represent temporal domain information

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 model training

**DISCRIMINATIVE POINT PROCESS MODELS**

**DPPM**

1. Collect correct/incorrect lattice arcs for each word type and compute the phone event point pattern for each word
2. Transform each arc into a fixed dimensional vector of event counts in uniform-time bins for each phone (1 duration)

**Discriminating Word Models**

1. Collect correct/incorrect lattice arcs for each word type and compute the phone event point pattern for each word
2. Transform each arc into a fixed dimensional vector of event counts in uniform-time bins for each phone (1 duration)
3. Use your favorite machine learning algorithm to train a classifier for each word (we use kernel machines and SVMs)

**EXPERIMENTS**

**Evaluation Corpus and Word Set**

- 430 hour HUB4+TDT4 broadcast news corpus, split into two equal parts for training and dev/eval

**Baseline: IBM Attila Recognizer**

- State-of-the-art acoustic model
- Quinphone states, 150K Gaussian (total), LDA, VTLN, RMMI, MLLR, iBMII

**Language model:** 4KBA 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**

**Discriminative PPM Scores:**

1. Train regularized least squares classifiers with radial basis function kernel learned for each confusable word (SVMS equivalent)
2. Training examples per word range from 2k to 25k
3. 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**

<table>
<thead>
<tr>
<th>Score Set</th>
<th>Lattice Only</th>
<th>Lattice + Forced Alignment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Attila</td>
<td>12.4%</td>
<td>17.3%</td>
</tr>
<tr>
<td>Discriminative PPM</td>
<td>16.1%</td>
<td>17.8%</td>
</tr>
<tr>
<td>Attila + DPPM</td>
<td>14.7%</td>
<td>16.4%</td>
</tr>
</tbody>
</table>

**Comparison:**

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