

# Modeling and Automatic Labeling of Hidden Events in Speech

*Elizabeth Shriberg and Andreas Stolcke*

*SRI International*

*Students:*

*Dilek Hakkani, Arthur Lu,  
Madelaine Plauche, Gokhan Tur*

*Collaborators:*

*Mari Ostendorf, Rebecca Bates*

# Introduction and Motivation

- Information on structural and functional units is crucial for understanding spoken language, e.g.:
  - disfluencies
  - sentence boundaries
  - topic boundaries
  - dialog acts
  
- View each unit as delimited by inter-word “events”.
  
- In written text, these events are either nonexistent (e.g., disfluencies) or explicit (e.g. punctuation).
  
- But in speech recognizer output — a stream of words — these events are “hidden”.

# Project Goals

- Automatically detect hidden events in ASR output
- Improve accuracy of ASR itself by modeling events

## Approach

- Exploit cues used by human listeners:
  - *What* is said (word recognition output)
  - *How* it is said (rhythm and melody = prosody)
- Combine standard ASR models with prosodic and language models for hidden events

# Overview of Talk

- Knowledge sources and statistical approach
- Results from three example tasks:
  - disfluency detection, sentence segmentation in conversational speech
  - topic segmentation in Broadcast News speech
  - event modeling for improved word recognition
- Progress to date and 1999 plans
- Conclusions

# Knowledge Sources

- From current ASR systems:
  - cepstrum-based acoustic model
  - word-based language model (LM)
  
- Added for event detection:
  - LMs relating words and hidden events, and constraints among events
  - prosodic model for hidden events
  
- All models are statistical, trained from data.
  
- Formal framework: find hidden events with highest posterior probability given words and prosody.

# Modeling Knowledge Sources

*A = acoustic features, W = words*

*E = events, F = prosodic features*

Estimate  $P(E|A,F)$  using:

- Recognizer acoustic model  $P(A|W)$
- Word language model  $P(W)$
- Event language model  $P(E,W)$
- Event prosodic model  $P(E|F,W)$

# Prosodic Modeling

- ❑ Speakers mark event boundaries prosodically.
- ❑ This information is present in acoustics, but not made explicit in standard ASR acoustic models.
- ❑ CART-style decision trees trained to predict events from features extracted around word boundaries.
- ❑ Duration: of pauses, phones, vowels, rhymes; normalized for segmental and speaker variation.
- ❑ Pitch: F0 patterns before, after, across boundary; distance from speaker baseline. Normalized, stylized and parameterized.

# Disfluency and Sentence Detection

- Switchboard (1.4M words, conversational speech)
- LM based on N-grams of hidden events and words
- Event types:

Sentence boundary	10.8%	he saw it twice * he liked it
Filled pause	2.9%	he uh * liked it
Repetition	1.9%	he * he liked it
False start	1.3%	it was * he liked it
Repair	1.2%	we * he liked it
ELSE	81.8%	he * liked it

# Results

- ❑ Used recognizer with 46.8% word error rate.  
Chance event detection error rate = 30.8%.
- ❑ LM with pauses: 26.2% (detection error)  
without pauses: 28.9%
- ❑ Prosody-only model: 27.1%
- ❑ Combined model: 25.1%
- ❑ Prosody more robust than LM to ASR errors.
- ❑ Prosodic features used: pause, duration, pitch.

# Topic Segmentation

- ❑ Broadcast News speech (50 hrs, 800K words for acoustic train/test; 130M words for LM train/test).
- ❑ LM: HMM with topic clusters as states, from unsupervised clustering (Dragon approach).
- ❑ Unigram LMs give topic likelihoods.
- ❑ Prosodic model gives topic-transition likelihoods.
- ❑ HMM search finds best overall segmentation.
- ❑ LM requires pre-segmenting into sentence-like units. Used pause threshold (also prosodic!)

# Results

- LM only: 20.7% (DARPA TDT weighted error)
- Prosody only: 17.8%
- Combined: 15.7% → 25% improvement
- Prosody alone does better than the words, and combined model gives substantial additional win.
- Prosody model trained on far smaller dataset.
- Prosodic features: pause (31% usage), pitch based on parameterized F0 model (45% usage).  
Main pitch feature: distance from speaker's floor.

# Improving Word Recognition: Proof of Concept Experiment

- ❑ Consider all possible event sequences for a given word hypothesis (disfluencies, sent. boundaries).
- ❑ Penalize word hypotheses whose likely event sequences are inconsistent with the prosody.
- ❑ Expected improvement small, since these events occur at only about 15% of word boundaries.
- ❑ Results on Switchboard: small but significant reduction in overall word error (1% absolute).
- ❑ Effect: prevent falsely hypothesized repeats, fillers.

# Project Progress

- 1997** - Disfluency/sentence boundary detection from known words
  - Dialog act classification on recognized words
  - Prosodic database for Switchboard
- 1998** - Event detection from recognized words
  - Robust pitch features for prosody model
  - Alternative classifier experiments (trees work well)
  - Model combination comparisons
  - Prosody and language model for Broadcast News, Topic segmentation on recognized words (supported by DARPA TRVS)
- 1999** - Event modeling for word recognition
  - Prosody for DARPA named entity task

# 1999 Plans

- Jointly model word and prosodic features
- Combine events across word hypotheses
- Enhance language models with parts of speech
- Use sentence detection for topic pre-segmentation
- Optimize prosodic features for word recognition

# Conclusions

- ❑ Detecting hidden events is crucial to language understanding.
- ❑ HMM-based language models detect events based on word information.
- ❑ Decision trees predict events from prosody.
- ❑ Prosodic information significantly improves disfluency detection, sentence boundary detection, topic segmentation, and word recognition.