

Markovian Combination of Language and Prosodic Models for Better Speech Understanding and Recognition

Andreas Stolcke Elizabeth Shriberg

Speech Technology and Research Laboratory

SRI International

Menlo Park, California

Acknowledgments & Collaborators

- Becky Bates and Mari Ostendorf (Boston U./U. of Washington)
- Dilek Hakkani-Tür and Gokhan Tür (Bilkent U./AT&T)
- Dan Jurafsky (U. of Colorado) & 1997 Johns Hopkins Summer Workshop
- Support by NSF, DARPA, NASA.

Motivation

- Most current work in speech and natural language processing is text-based.
- Spoken language has cues not present in written language: **prosody**.
- Linguistics literature describes how speakers use prosody systematically to convey semantic and pragmatic structure (e.g., boundary tones as indicators of sentence ends).
- Recognizing more than just words is increasingly important in applications (e.g., automatic punctuation).
- Prosodic cues may also increase robustness to recognition errors compared to models based only on words.

Challenges for Computational Prosody

- What cues are useful?
- How can they be automatically extracted?
- What computational models are appropriate for prosody?
- How should prosodic information be combined with other (especially lexical) cues?

In this talk:

Focus on **integration of lexical and prosodic** models using **hidden Markov models (HMMs)**.

Overview

- Elements of our prosody modeling approach
 - Modeling philosophy
 - Prosodic features
 - Prosodic models
 - Integration with lexical models
- Applications (see abstract references for others)
 - Sentence segmentation
 - Topic segmentation
 - Dialog act classification
 - Word recognition
- Conclusions and Outlook

Part I: Modeling Framework

Modeling Philosophy

- **Direct modeling:** models represent measurable correlates of prosody; no intermediate phonological categories that require human judgment (e.g., no tone labels) [cf. Batliner et al., Eurospeech 2001]
- **Automatic extraction of prosodic features:** based on automatic recognition, pitch tracking and regularization, energy.
- **Integration with lexical models:** construct computationally efficient joint models of prosodic and lexical observations.
- **Statistical modeling:** all models are probabilistic and can be configured for different tasks (e.g., structure tagging or word recognition)

Prosodic Features

- Duration features (pause, vowel/coda/syllable lengths) determined by automatic phone-level alignment of words (true or recognized)
 - normalized for speaker and speaking rate
- Pitch features
 - maxima/minima/averages over defined regions: whole utterance, word before/after location of interest, fixed-length window before/after location of interest.
 - slopes and differences
- Energy features
- Related non-prosodic features: gender, speaker changes, relative positions
- Features are designed to be application-dependent!

Prosodic Models

- **Decision trees** estimate posterior probability of target classes from prosodic features.
- Important: DTs allow inspection of feature use for understanding and debugging.
- **Feature subset selection** wrapper algorithm counteracts sub-optimal greediness of CART tree growing.
- **Downsampling** of training set creates equal priors of target classes
 - Forces DT to model infrequent classes.
 - Facilitates integration with language models.

Integration with Lexical Models

- Lexical cues (words and phrases) complement prosody
- Good performance requires both knowledge sources to be integrated.
- Direct use of lexical cues alongside prosodic features is difficult due to large feature space.
- Our approach: model relation between target classes and words using statistical language models.
- Combine prosodic and language models using various independence assumptions.
- Use transcribed or recognized words.
- No changes to recognizer acoustic models.

Probabilistic Modeling

- **Ingredients:** Structure classes S ; prosodic features F . words W
Example: Sentence segmentation
 - classifies each word boundary
 - $S = \{\text{s-final}, \text{s-internal}\}$
 - $F =$ features extracted from window around each word boundary
 - $W =$ recognized/transcribed word string
- **Prosodic classifier:** estimates $P_{DT}(S|F, W)$ where W includes only alignment, not word identities.
- **Language model:** estimates $P_{LM}(W, S)$, thus $P_{LM}(S|W) = \frac{P(W, S)}{P(W)}$
- **Goal:** estimate $P(S|W, F)$ using both LM and prosody.

Model with HMMs

- Idea: put prosodic observations into the LM.
- Build joint model $P(W, S, F)$, estimate $P(S|W, F) = \frac{P(W, S, F)}{\sum_S P(W, S, F)}$.
- For efficiency, represent $P(W, S, F)$ as a HMM, with states encoding S .
- HMM structure is task-dependent.
- Use decision tree posteriors as observation pseudo-likelihoods.
Note: assumes equal priors for all S .
- Word-based and prosodic likelihoods are weighted for best results (optimized on held-out data).
- Good: joint model; single score combination weight to optimize.
- Bad: independence assumptions in HMM.

Part II: Applications

Application 1: Sentence Segmentation

Task

- Classify each word boundary into sentence/no-sentence boundary
- Important for usable speech recognition output (automatic punctuation)
- Compare Broadcast News and Switchboard speaking styles

Prosodic Features

- Extracted from windowed regions around each word boundary

Sentence Segmentation

Modeling

- 4-gram language model trained on text with sentence boundary tokens:

... in washington <s> there is ...

- HMM formulation:
 - the states denote the 4-grams of words and sentence boundaries.
 - transition probabilities were obtained from the LM.
 - prosodic observation likelihoods (from DT) attached to states.
 - forward-backward algorithm computes posterior probabilities of boundary types at each location: $P(S|F, W)$
- Both LM and prosodic model are conditioned on speaker changes.

BN Sentence Segmentation Results

- Chance classifier outputs “no boundary” everywhere.
- Lower error bound due to inserted/deleted word boundaries in recognition.
- Word recognition error rate $\approx 30\%$.

| Model | Classification Error (%) | |
|---------------------------|--------------------------|------------------|
| | True Words | Recognized Words |
| Chance | 6.2 | 13.3 |
| LM only (700K words) | 4.8 | 12.4 |
| LM only (130M words) | 4.1 | 11.8 |
| Prosody only (700K words) | 3.6 | 10.9 |
| Combined | 3.3 | 10.8 |
| Lower bound | 0.0 | 7.9 |

Switchboard Sentence Segmentation Results

| Model | Classification Error (%) | |
|--------------|--------------------------|------------------|
| | True Words | Recognized Words |
| Chance | 11.0 | 25.8 |
| LM only | 4.3 | 22.8 |
| Prosody only | 6.7 | 22.9 |
| Combined | 4.0 | 22.2 |
| Lower bound | 0.0 | 17.6 |

Word recognition error rate $\approx 47\%$.

Application 2: Topic Segmentation

Task

- Classify word boundaries in Broadcast News shows as topic changes/no topic changes.
- Many applications in audio indexing, information extraction, etc.

Prosodic features

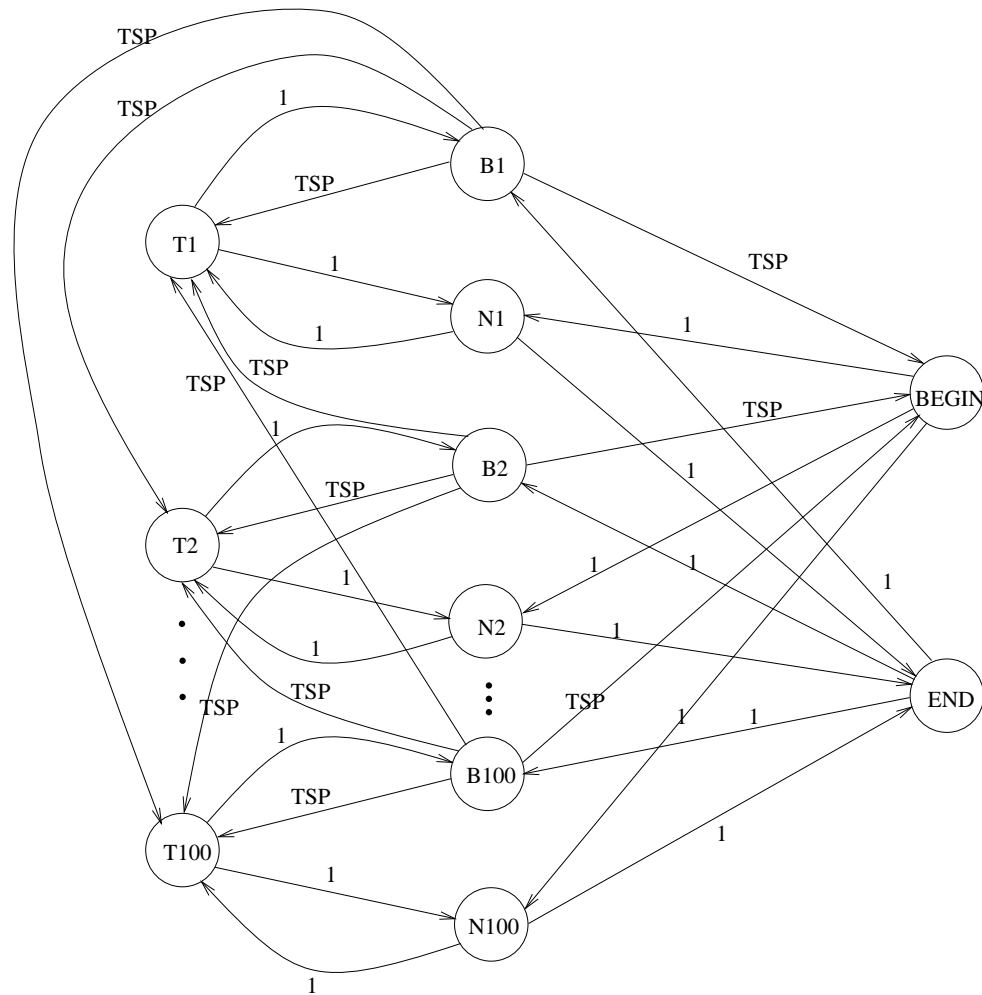
- Pause threshold used to presegment data into pseudo-sentences
- Features extracted around potential topic boundaries used in model

Topic Segmentation

Modeling

- Topic segmenter based on a model developed by Dragon Systems.
- Based on 100 individual unigram topic language models, obtained by automatic clustering.
- HMM states correspond to topic clusters and emit sentences using unigram likelihoods.
- 2 extra states for topic-initial and final sentences.
- special states represent potential topic boundaries.
- prosodic feature classifier estimates boundary state likelihoods
- between-topic transitions are penalized by a *topic switch penalty* (TSP).
- Viterbi algorithm finds best segmentation.

HMM Topology



Topic Segmentation Results

| Segmentation Cost | | |
|-------------------|------------|------------------|
| Model | True Words | Recognized Words |
| Chance | 0.3 | 0.3 |
| LM only | 0.1895 | 0.1897 |
| Prosody only | 0.1657 | 0.1731 |
| Combined | 0.1377 | 0.1438 |

TDT2 weighted error metric:

$$\text{Segmentation cost} = (\text{False Alarm} \times 0.7) + (\text{Miss} \times 0.3)$$

Application 3: Dialog Act Tagging

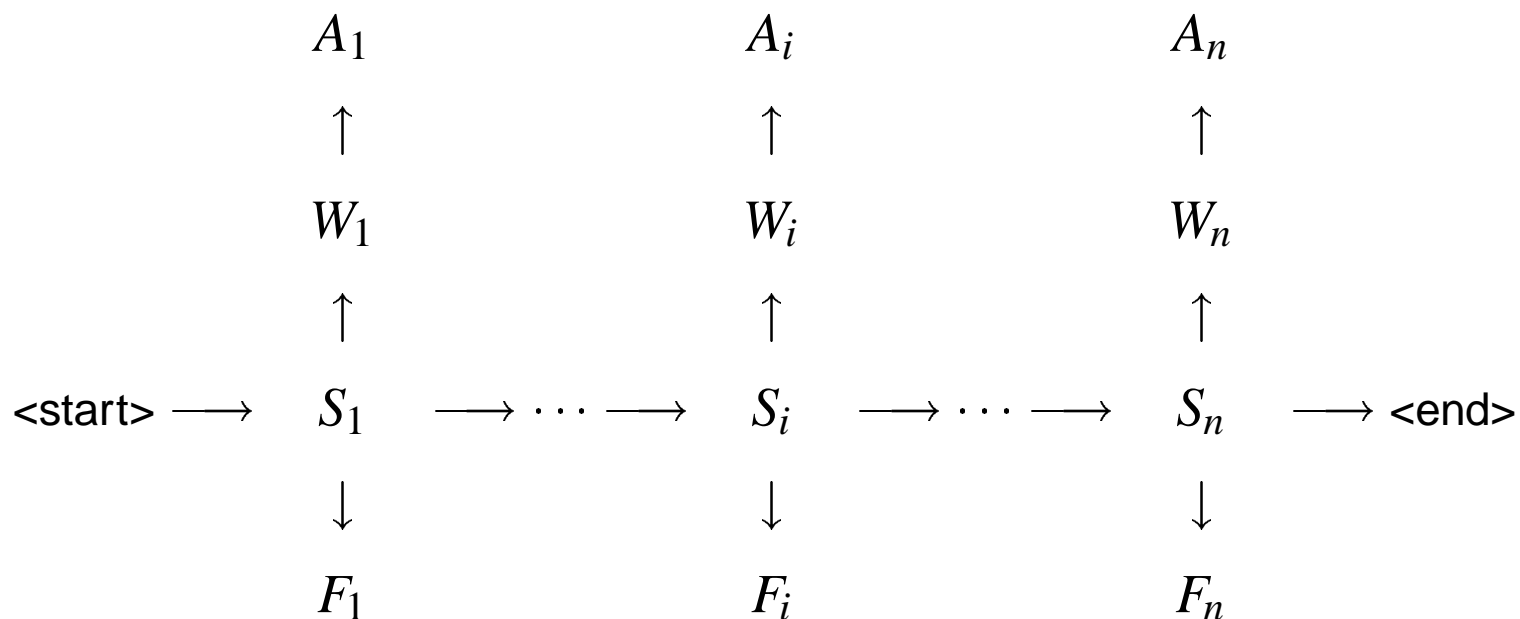
Task

- Classify whole utterances in spontaneous conversations into *dialog act* types
- DA types: statement, agreement, backchannel, question, . . .
- Assumes dialog act boundaries are known (but cf. Warnke et al. 1997)

Prosodic features

- extracted from whole utterances
- can help dialog classification
 - as additional knowledge source
 - to distinguish between lexically ambiguous DAs
- Example: Statement (“They’re gone!”) vs. Question (“They’re gone?”).

Graphical Model for DA Tagging



$P(F_i|S_i)$ prosodic decision tree

$P(W_i|S_i)$ dialog act language models

$P(S_i|S_{i-1})$ dialog act N-gram

$P(A_i|S_i)$ speech recognizer

Overall DA Tagging Results

Dialog act tagging error (%) based on recognized words (WER = 41%):

| Model | Dialog Grammar | | |
|--------------|----------------|---------|--------|
| | None | Unigram | Bigram |
| Chance | | 65.0 | |
| LM only | 57.2 | 38.2 | 35.7 |
| Prosody only | 61.1 | 51.7 | 50.3 |
| Combined | 43.5 | 37.6 | 35.0 |

- 42 dialog act types.
- Prosody gives small but significant gain over LM alone.
- Win is small due to skew in DA distribution.
- Prosody more effective in discriminating specific DA types.

Task 4: Word Recognition

- Can we benefit from existing prosodic classifier in word recognition?
- Prosody is not currently used in state-of-the-art large vocabulary speech recognizers.
- Several approaches have been tried, including:
 - Modeling prosodic aspects of words themselves (e.g., word and phone durations)
 - Modeling prosodic properties of linguistic structure *above* the word level.
- Examples of second approach:
 - Prosodic scoring of parse structures (Veilleux & Ostendorf 1993)
 - Conditioning on dialog acts (JHU Workshop 1997, Taylor et al. 1998)

Prosodic Hidden Event Modeling for Recognition

Model sentence boundaries and disfluencies: “**hidden events**”

- **Idea:** Score agreement of words, hidden events and prosodic correlates.
- **Example:** *he* <REP> *he liked it* (disfluent repetition)
Prosodic features: First “he” is typically lengthened and followed by a pause.
- Standard speech recognizer: $P(W|A) \propto P(W)P(A|W)$
 A = recognizer front-end features, $P(W)$ = language model.
- New approach: $P(W|A, F) \propto P(W, F)P(A|W)$
 $P(W, F) = \sum_S P(W, S, F)$ from hidden event HMM.
- Same HMM and prosodic model as used for hidden event detection!
- **Note:** Expected win small: 18% of boundaries have non-null hidden event.

Word Recognition Results

| Model | WER | Sub | Del | Ins |
|-------------------------|------|------|------|-----|
| Standard N-gram | 47.9 | 31.1 | 12.2 | 4.6 |
| HE N-gram, no prosody | 47.6 | 30.4 | 13.3 | 3.9 |
| HE N-gram, with prosody | 47.0 | 29.7 | 14.1 | 3.2 |

- Error reduction small, but highly consistent ($p < 10^{-6}$).
- Hidden event language model alone (without prosody) gives small win (-03%, $p < 0.02$).
- Overall error reduction is due to reduced substitutions and insertions.
- Anecdotal observation: Prosody reduces false alarms of frequent words likely to occur at sentence starts and hesitations (“I”, “and”, “the”).

Conclusions

- Prosody can enhance a number of speech tagging tasks.
- Positive results for sentence segmentation, disfluency detection, topic segmentation, and dialog act tagging.
- Prosodic information can improve accuracy of word-only models significantly.
- Prosody can be leveraged indirectly for word recognition by modeling events between words: sentence boundaries, hesitations, repairs, etc.
- HMMs (based on statistical language models) provide efficient and effective framework for integrating word-based and prosodic models.

Thank You!