

Prosody Modeling for Speech Recognition and Understanding

Ramana Rao Gadde Elizabeth Shriberg Andreas Stolcke

Dilek Hakkani-Tür Gökhan Tür

Speech Technology and Research Laboratory

SRI International

Menlo Park, California

Hub-5 Conversational Speech Understanding Workshop

June 10, 1999

Why Prosody for LVCSR?

- Speech prosody = duration, pitch, energy, especially above the segment level
- Prosody is extremely important to our perception of natural speech (e.g., witness role in speech synthesis)
- Prosody is not captured in today's LVCSR systems

Goals

- How can prosody help in improving word recognition?
- How can prosody help in speech understanding (recognizing more than words)?

Outline

- Prosody for word recognition I: Gaussian mixture modeling of word durations
- Prosody for detecting “Hidden Events”: disfluency detection, sentence segmentation, topic segmentation
- Prosody for word recognition II: Modeling hidden events can lower word error
- Conclusions

Duration Modeling With Gaussian Mixtures

- Suprasegmentals in ASR
- Word level modeling of Suprasegmentals
- Modeling duration with Gaussian Mixtures
- Experiments
- Future Work

Suprasegmentals in ASR

- Current ASR systems model mostly, segmental features like cepstra. Humans are known to use suprasegmental features like duration and pitch.
- Several studies showed that suprasegmental features are related to the structure of the sentence. Many models were proposed to exploit this in ASR systems (Veilleux and Ostendorf, ICASSP'93; Hunt, Aust.JCAI'94).
- In another study, it was shown that word models incorporating lexical stress can perform better than models without stress (Hieronymus, McKelvie and McInnes, ICASSP'92).

- Word level F_0 patterns were shown to be useful for detecting word boundaries in Indian languages and German (Ramana Rao and Srichand, ICSLP'96).
- Modeling prosody led to significant reduction in errors for connected number recognition in Spanish (Lopez, Caminero, Cortazar and Hernandez, ICASSP'99).

Word level modeling of Suprasegmentals

- Most research on modeling word level suprasegmentals in English is based on modeling lexical stress.
- Stress is a perceptual feature and is not a measurable quantity. Its acoustic correlates are syllable duration, pitch, energy and vowel quality.
- In absence of a well defined relation between the acoustic features and stress, it is difficult to identify stress. This means that it is difficult to automatically train models of stress from training data.

Modeling Duration with Gaussian Mixtures

- Our aim is to develop models to represent the suprasegmental information in speech in a way suitable to current SR systems. The models should be trainable in an automatic way and they should easily fit into the ASR process.
- We propose to model the suprasegmental features of duration, pitch and energy at the word level.
- Each word is represented by a vector of the features. For example, the word "that", pronounced as "dh+ae+t", may be represented by a 3-dimensional durational feature vector (10.0 8.0 4.0), where the three values, 10.0, 8.0 and 4.0 represent the durations of the three phones, "dh", "ae", and "t" respectively.

- This representation allows us to model a word using suprasegmental features by models such as GMMs which can be trained from the training data.
- During recognition, we can compute the feature vector for the words in the hypotheses, which can be scored by the suprasegmental models. This score can then be combined with the acoustic and language model scores.

Experiments

- We trained acoustic models using LVCSR training data containing nearly 125K utterances.
- Duration models were also trained using this training set.
- For testing, we used the male subset of the 96 English eval set. It had 20 speakers and 1301 utterances.
- For each utterance in the test set, we generated 100-best hypotheses. These were scored using the prosodic model.
- The prosodic score was weighted and added to the acoustic+language score and the hypothesis with the best score was chosen.

Results

The results are shown for the full test set (Full) and for a subset of 466 utterances (Sub_466) whose 100-best contained the correct hyp.

Model	Test set	
	Full	Sub_466
Baseline	59.8%	34.5%
+prosody	59.9%	33.7%

Future Work

- Investigate automatic duration model generation to handle OOVs and fast speech.
- Investigate optimal feature set.
- Investigate if the prosodic models can be used across different acoustic conditions (wide band/narrow band, clean/noisy etc.).

Hidden Events: Recognizing more than words

- LVCSR systems output a string of words
- Speech understanding also needs *hidden events*

I'm sure <S> I mean, I <REP> I don't know

- Disfluencies (hesitations, restarts, repairs)
- Sentence boundaries
- Topic boundaries
- Linguistic literature indicates prosody is important cue for these events
- Prosody can partly compensate for lack of textual cues (punctuation, case, paragraphs, headings) in speech

Hidden Event Types

Six event types

Event class	Tag	Freq.	Example
Sentence boundary	S	10.8%	I haven't seen it * Not sure I like it
Filled pause	FP	2.9%	he uh * liked it
Repetition	REP	1.9%	he * he liked it
Deletion	DEL	1.3%	it was * he liked it
Repair	OthDF	1.2%	he * she liked it
Else/fluent	else	81.8%	she * liked it

Models

What is modeled

- Recognizer acoustic features A
- Word string $W = W_1 W_2 \dots W_n$
- NEW: Prosodic features F
- NEW: Inter-word events $E = E_1 E_2 \dots E_n$

Models used to estimate $P(E|A, F)$

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

Prosodic Modeling

- CART-style decision trees estimate $P(E_i|F_i, W)$
- Use only word/phone alignment, not word identity (robust to recognition errors)
- Models are trained on forced alignments, tested on imperfect recognition hypotheses
- Features used:
 - Duration: of pauses, final vowels and rhymes, normalized for phone durations and for speaker
 - Pitch: F0 patterns; before, after, and across boundary; change relative to estimated speaker baseline. Pitch tracks postprocessed for regularization; octave error modeling
- Model acoustic measurements directly, without phonological labels.

Event Language Model

- N-gram model estimates probability of joint word/event sequence: $P(W, E)$
- Trained on annotated transcript:

Right <S> I <REP> I don't I'm not sure . . .

- Testing as hidden event HMM.
Forward-backward algorithm computes $P(E_i|W)$
- Two versions
 - “Seg N-gram”: includes turn and pause markers
 - “No-seg N-gram”: does not include turns and pauses

Note: segmentation cues are also part of prosodic model

Data and Method

- 1100 Switchboard conversations
 - 1.2M words for training
 - 231K words for testing using known words
 - 18K words for testing on recognized words
- LDC transcripts hand-labeled for events
- Built parallel database of knowledge sources
 - word transcripts
 - event annotations
 - phone-level alignments from recognizer
 - frame- and word-level acoustic measurements

Hidden Event Recognition Results

Event error rates using individual knowledge sources (WER = 46.8%)

Model Type	Known Words	Recognized words
Chance	18.2	30.8
Prosodic Tree	11.1	27.1
No-seg N-gram	10.0	28.9
Seg N-gram	7.3	26.2

- N-gram (with turns and pauses) is single best detection model
- Substantial degradation on recognized words (but less than predicted by word error rate)
- Prosody model degrades less than N-gram (as expected)

Event Confusion Matrix

Lowest accuracies for false starts and repairs (lowest frequency events)

		Detected Events					
		S	else	FP	DEL	OthDF	REP
True Events	S	16880	5065	0	111	31	33
	else	3594	162847	0	439	174	170
	FP	0	0	5879	0	0	0
	DEL	393	1524	0	660	131	48
	OthDF	218	1338	0	191	442	341
	REP	43	892	0	34	97	2856

Results Combining Words and Prosody

Model Type	Known Words	Recognized words
Seg N-gram	7.3	26.2
Interpolated	7.1	25.1
Joint Tree	6.9	26.7

- Prosody gives significant win over word-based model
- Interpolation of posteriors gives 4% error reduction over N-gram
- Joint tree likely suffers from train/test mismatch: needs to be trained on recognized words
- Other combination methods under investigation

Sentence Segmentation Results

Special case of hidden event recognition (results on true words)

Model	Switchboard	Broadcast News	
Chance	11.0	6.1	
Prosody only	6.77	4.1	
LM only	4.64	4.8	4.0
Prosody + LM	4.45	3.6	3.2
		700K LM	130M LM

- “700K LM” trained on same amount of data as prosody model
“130M LM” trained on 1996 Broadcast News corpus
- Significant wins from combining words and prosody
- LM modeling *all* hidden events (including DFs) is best for sentence segmentation

Feature Usage

Results from automatic feature subset selection algorithm:

Percentages reflect relative number of times a feature is queried in tree.

Switchboard

- phones and rhyme duration (44.5%)
- pause duration (33.2%)
- turn boundaries (22.3%)
- No F0 features!

Broadcast News

- pause duration (49%)
- F0 difference (23%)
- turn boundaries (23%)
- rhyme duration (4%)

Topic Segmentation

- Task: Find topic boundaries in Broadcast News shows
- Word-based Dragon topic segmenter with SRI improvements
- Language Model: HMM with states corresponding to topic clusters (from unsupervised clustering)
- Each HMM state emits one sentence (or chunk of words delimited by pauses); computes topic likelihoods using unigram LM
- Topic transitions have additional likelihoods from prosodic decision tree.

Results

Model Type	Known Words % weighted error	Recognized words % weighted error
Word-based	19.5	20.9
Prosodic DT	18.9	19.9
Combined	15.0	16.8

- Prosody alone does better than words on this task
- Significant wins from model combination (-20%)
- Pre-segmentation based on pauses is better than using true sentence boundaries

Feature Usage

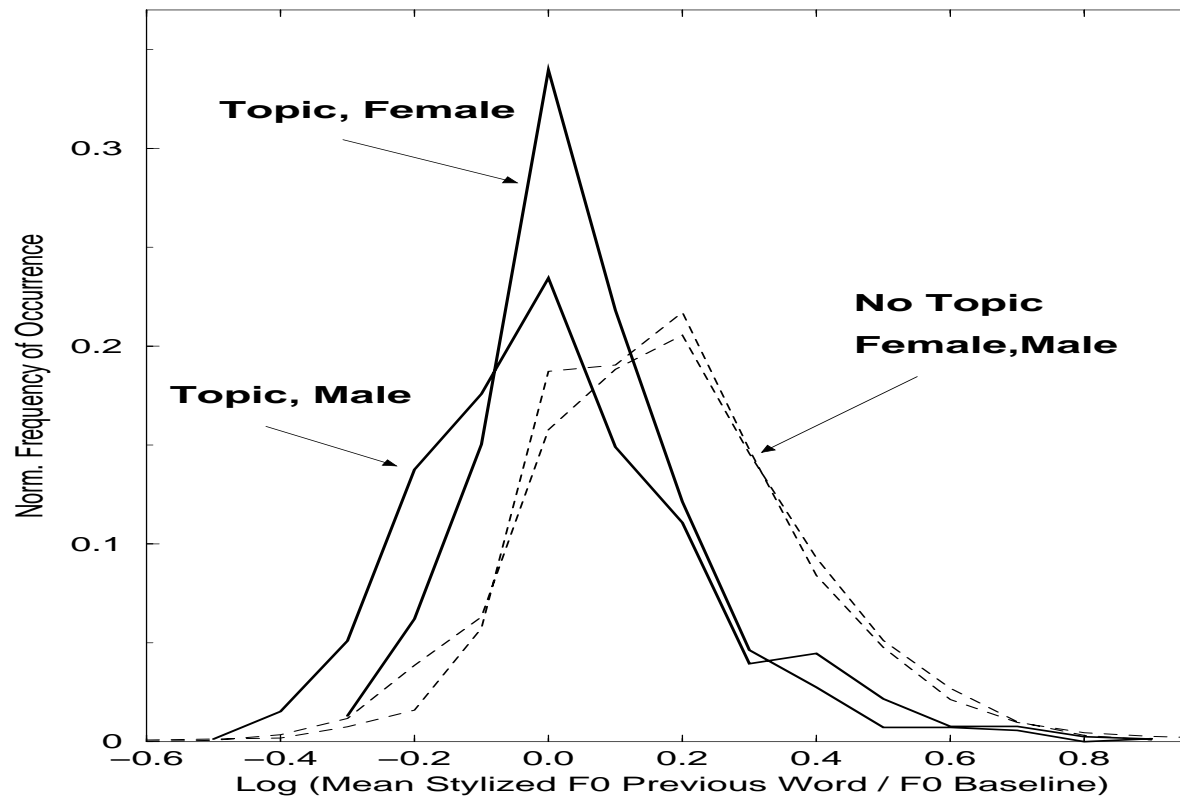
Relative feature usage (on Broadcast News)

- pause duration (30%)
- F0 relative to baseline (28%)
- gender (17%)
- turn boundary (16%)
- distance into story or turn (9%)

Why is gender being used? Inspection shows not due to normalization, but rather stylistic differences between M/F in signaling topic boundaries intonationally.

Gender Differences

M/F similar for no-topic, F closer to their baselines than M for topic:



Hidden Event Modeling for Improved Word Recognition?

Intuition: Penalize word hypotheses that entail hidden events that are not supported by prosody

Formally: Condition language model on prosody via hidden events

$$\begin{aligned}
 P(W|A, F) &= \frac{P(W|F)P(A|W, F)}{P(A|F)} \\
 &\propto P(W|F)P(A|W) \\
 &\propto P(W, F)P(A|W) \\
 &= \sum_E P(W, E, F)P(A|W)
 \end{aligned}$$

$\sum_E P(W, E, F)$ can be computed by the hidden event HMM with prosodic likelihoods (from decision tree)

Note: Expected win is small; only 18% of word boundaries have a hidden event.

Results

- Switchboard corpus (900 conversations for training, 6 for tuning, 19 for testing)
- Rescored 100-best lists generated with standard bigram recognizer
- Use same prosodic DT as in hidden event detection

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 ($p < 10^{-6}$).
- Overall reduction due to reduced substitutions and insertions.

Analysis

- **Overall observation:** Prosody reduces false alarms of frequent words likely to occur at sentence starts and hesitations (“I”, “and”, “the”).
- **Error count differences**

Baseline	HE model	Ref. words	Error Δ
correct	correct	11402	0
incorrect	incorrect	8042	-118
correct	incorrect	597	+606
incorrect	correct	569	-619

Examples

Sentence boundaries constrain words

(2131-B-0053) ... that at church **to** <S>

→ ... that at church **too** <S>

Disfluencies constrain words

(2461-B-0044) ... to really hurt **to** <REP> the middle class

→ ... to really hurt **the** <REP> the middle class

Improved filled pause recognition

(3528-B-0038) ... to perform in **and** cold weather

→ ... to perform in **UH** cold weather

Reduced false DF recognition

(2753-A-0008) ... <S> **the** the source of ...

→ ... <S> **but** the source of ...

Conclusions

- Word duration modeling using gaussian mixtures is promising, but needs more work.
- Disfluency and sentence boundary detection in conversational speech: encouraging results with combined prosody and language models, mostly based on duration features.
- Sentence and topic segmentation: prosody (including F0) is strong cue on Broadcast News.
- Consistent error reduction in LVCSR word error by conditioning language model on the prosody evidence for hidden events.