

Combining Words and Prosody in the Language Model

Andreas Stolcke

Speech Technology and Research Laboratory

SRI International

Menlo Park, California

IMA Workshop

November 1, 2000

Why Prosody?

- 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
- Prosody correlates with lexical, syntactic, and semantic structure.
- Prosody helps disambiguate meaning:

No linguists are helpful.

No! Linguists are helpful.

- Cf. recent work on prosody for syntactic disambiguation and speech translation [Verbmobil]

Hidden Events

- Model various kinds of *word boundaries* in the word stream.

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

- Linguistic literature indicates prosody is important cue for these events
- Original motivation: need to detect sentence boundaries and disfluencies for parsing, interpretation of spoken language.
- Here: Use hidden events to
 - build a more detailed language model
 - tie the LM to prosodic cues
- Intuition: penalize word hypotheses whose prosody is not consistent with hidden events.

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

Random Variables and 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 $E = E_1 E_2 \dots E_n$

Component models

- 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 (robustness to errors)
- Train on true words, test on errorful 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.
- Model acoustic measurements directly, without phonological labels.
- By sampling the training set we can get scaled likelihoods:

$$P(F_i|E_i, W) = C \cdot P(E_i|F_i, W)$$

Event Language Model

- N-gram model estimates probability of joint word/event sequence: $P(W, E)$
- **Training:** supervised, from annotated transcripts:

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

- **Evaluation:** Marginalize over all possible hidden events, using dynamic programming (equivalent to a higher-order HMM).

Conditioning the Language Model on Prosody

- 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)
- Treats prosodic features as conditionally independent given events and words:

$$P(W, E, F) = P(W, E) \prod_i P(F_i|E_i, W)$$

Experiments

- Switchboard corpus (900 conversations for training, 6 for tuning, 19 for testing)
- Rescored 100-best lists generated with standard bigram recognizer
- Prosodic likelihoods estimated by decision tree, from durational features only (syllable lengthening, pauses)
- **Note:** Expected win is small; only 18% of word boundaries have a hidden event.

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

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 ...

Summary

- Let's not forget to look at new knowledge source within the speech signal
- Prosody is promising, can be tied fairly directly to word-level models.
- Currently using “low-tech” N-gram and HMM techniques for model combination. Can we do better?
- What is best way to model interactions between discrete, word-related events and continuous-valued prosodic features?