Speech Recognition Models of the Interdependence Among Prosody, Syntax, and Segmental Acoustics

Mark Hasegawa-Johnson
jhasegaw@uiuc.edu

Jennifer Cole, Chilin Shih, Ken Chen, Aaron Cohen, Sandra Chavarria, Heejin Kim, Taejin Yoon, Sarah Borys, and Jeung-Yoon Choi
Outline

• Prosodic Constraints for ASR
• Prosodically Transcribed Speech Corpus
• Prosody-dependent speech recognition
  – Framework
  – How Prosody Reduces Word Error Rate
• Acoustic models
  – Factored prosody-dependent allophones
  – Knowledge-based factoring: pitch & duration
  – Allophone clustering: spectral envelope
• Language models
  – Factored syntactic-prosodic N-gram
  – Syntactic correlates of prosody
Prosodic Constraints for ASR

Goals

• Disambiguate sentences with similar phonemic content.
• Create speech recognition algorithms which will fail less often in noisy environments.

Example

“The nurse brought a big Ernie doll.”
“The nurse brought a bigger needle.”
Prosodic Constraints for ASR

What is Prosody?
Why is Prosody Useful?
Why is Prosody Ignored by ASR?
What Can We Do About It?
What is Prosody?

Lexical Stress (Phonological):
- Lexical Stress is marked in the dictionary.
- Perceptual Correlates: stressed syllable may receive prominence.

Phrasing and Prominence (Perceptual):
- Phrasing and Prominence are controlled by the speaker to suggest the correct syntactic and pragmatic parse of a sentence.
- Acoustic Correlates: pitch, duration, glottalization, energy, and spectral envelope.
Perceptual and Acoustic Correlates of Stress

Phonological Domain

Perceptual Domain

Acoustic Domain

Stressed

Phrasal Prominence

P1

Energy

P2

Minor Prominence

F0

1-P1-P2

No Prominence

Duration

Unstressed

Glottalization

No Prominence

Spectral Envelope

1-P1-P2
What is Prosody?

Prosody is a System of Constraints:

• Syntax and semantics constrain $p(w_2|w_1)$
• Prosody constrains $p(O|W)$

Prosody is Hierarchical and Non-Local:

• Phrase-final lengthening and phrase-initial glottalization increase with boundary depth
• Location of prominences is constrained by phrase structure
Syntactic Hierarchy
(Perhaps Semantically Tagged)

Constrains the probability of any particular phone sequence.

Prosodic Hierarchy

Constrains the implementation of the phone sequence.

Sentence

Subject

Pronoun

We

/w/ /i/ /v/ /E/ /k/ ....

Predicate

Transitive

Dir. Obj.

Mod. Adj. Noun

a

nice

beach.

Sentence

Breath Group

Intonational Phrase

Minor Group

Word

We

/w/ /i/ /v/ /E/ /k/ ....

wreck

a

nice

beach.

Word

Onset Rhyme

Syllable

Onset

Rhyme

....
Why is Prosody Useful?

1. Humans extremely sensitive to prosody
   • Infants use prosody to learn new vocabulary (Jucszyk, 1989).

2. Prosody is audible in noise
   • Low-frequency acoustic correlates (energy, F0)

3. Prosody disambiguates confusable words
   • Experiment: destroy all fine phonetic information, keep only 6 manner classes.
     Average cohort size = 5.0 (std=19.6, max=538)
   • Keep manner classes, plus lexical stress.
     Average cohort size = 3.4 (std=11.6, max=333)
Prosody modeled in our system

- Two binary tag variables (Toneless ToBI):
  - The Pitch Accent (*)
  - The Intonational Phrase Boundary (%)
- Both are highly correlated with acoustics and syntax.
  - Pitch accents: pitch excursion (H*, L*); encode syntax information (e.g. content/function word distinction).
  - IPBs: preboundary lengthening, boundary tones, pause, etc.; Highly correlated with syntactic phrase boundaries
“Toneless ToBI” Prosodic Transcription

• Tagged Transcription:
  Wanted*% chief* justice* of the Massachusetts* supreme court*%
  – % is an intonational phrase boundary
  – * denotes pitch accented word

• Lexicon:
  – Each word has four entries
    • wanted, wanted*, wanted%, wanted*%
  – IP boundary applies to phones in rhyme of final syllable
    • wanted% w aa n t ax% d%
  – Accent applies to phones in lexically stressed syllable
    • wanted* w* aa* n* t ax d
The Corpus

- The Boston University Radio News Corpus
  - Stories read by 7 professional radio announcers
  - 5k vocabulary
  - 25k word tokens
  - 3 hours clean speech
  - No disfluency
  - Expressive and well-behaved prosody
- 85% utterances are selected randomly as training, 5% for development-test and the rest 10% for testing.
- Small by ASR standards, but is the largest ToBI-transcribed English corpus
Example: “(if they think they can drink and drive, and) get away with it, they’ll pay.”
Why is Prosody Ignored by ASR?

The search problem:
- Prosodic constraints are non-local, and are therefore difficult to use in an efficient search algorithm.

1. The normalization problem:
- Acoustic features must be normalized to account for speaker variability
The search problem: Prosody dependent speech recognition

\[
\hat{W} = \arg \max_{W} p(O \mid Q, H) \\
\cdot p(Q, H \mid W, P) \\
\cdot p(W, P)
\]

• Advantages:
  – A natural extension of PI-ASR
  – Allow the convenient integration of useful linguistic knowledge at different levels
  – Flexible
A Bayesian network view of a speech utterance

- **X**: acoustic-phonetic observations
- **Y**: acoustic-prosodic observations
- **Q**: phonemes
- **H**: phone-level prosodic tags
- **W**: words
- **P**: word-level prosodic tags
- **S**: syntax
- **M**: message
Prosodic tags as “hidden speaking mode” variables
(inspired by Ostendorf et al., 1996, Stolcke et al., 1999)

\[ W = \arg\max_W \max_{Q, A, B} p(X, Y | Q, A, B) \, p(Q, A, B | W, S, P) \, p(W, S, P) \]

<table>
<thead>
<tr>
<th>Standard Variable</th>
<th>Hidden Speaking Mode</th>
<th>Gloss</th>
</tr>
</thead>
</table>
| Word              | \(W=[w_1,\ldots,w_M]\) | \(P=[p_1,\ldots,p_M],\)  
|                   |                      | \(S=[s_1,\ldots,s_M]\)  
|                   |                      | Prosodic tags, Syntactic tags |
| Allophone         | \(Q=[q_1,\ldots,q_L]\) | \(A=[a_1,\ldots,a_L],\)  
|                   |                      | \(B=[b_1,\ldots,b_L]\)  
|                   |                      | Accented phone, Boundary phone |
| Acoustic Features | \(X=[x_1,\ldots,x_T]\) | \(Y=[y_1,\ldots,y_T]\)  
|                   |                      | F0 observations |
Prosody dependent language modeling

\[ p(w_i|w_{i-1}) \Rightarrow p(w_ip_i|w_{i-1},p_{i-1}) \]

Prosodically tagged words:

\textit{cats*} climb \textit{trees*%}

Prosody and word string jointly modeled:

\[ p(\text{trees*%} | \text{cats* climb}) \]
Prosody dependent pronunciation modeling

\[ p(Q_i|w_i) \Rightarrow p(Q_i,H_i|w_i,p_i) \]

1. Phrasal pitch accent affects phones in lexically stressed syllable
   
   *above* \ ax b ah v
   *above* \ ax b ah v*

2. IP boundary affects phones in phrase-final rhyme
   
   *above*% \ ax b ah% v%
   *above*% \ ax b ah% v*%
Prosody dependent acoustic modeling

- Prosody dependent allophone models
  \[ \Lambda(q) \Rightarrow \Lambda(q,h) \]:
  - Acoustic-phonetic observation PDF
    \[ b(X|q) \Rightarrow b(X|q,h) \]
  - Duration PMF
    \[ d(q) \Rightarrow d(q,h) \]
  - Acoustic-prosodic observation PDF
    \[ f(Y|q,h) \]
How Prosody Improves Word Recognition

- Discriminant function, prosody-independent
  - \( W_T \) = true word sequence
  - \( W_i \) = competing false word sequence
  - \( O \) = sequence of acoustic spectra

\[
\Phi(W_T; O) = E_{W_T,O} \{ \log p(W_T|O) \} \\
= - E_{W_T,O} \{ \log ( \sum_i \eta_i ) \}
\]

\[
\eta_i = \frac{p(O|W_i)}{p(O|W_T)} \times \frac{p(W_i)}{p(W_T)}
\]
How Prosody Improves Word Recognition

- Discriminant function, prosody-dependent
  - $P_T =$ True prosody
  - $P_i =$ Optimum prosody for false word sequence $W_i$

$$
\Phi_P(W_T;O) = \mathbb{E}_{W_T,O} \{ \log p'(W_T|O) \} = - \mathbb{E}_{W_T,O} \{ \log (\Sigma \eta'_i) \}
$$

$$
\eta'_i = \frac{p(O|W_i,P_i)}{p(O|W_T,P_T)} \times \frac{p(W_i,P_i)}{p(W_T,P_T)}
$$
How Prosody Improves Word Recognition

- Acoustically likely prosody must be…
- unlikely to co-occur with…
- an acoustically likely incorrect word string…
- most of the time.

\[ \Phi_P(W_T;O) > \Phi(W_T;O) \]

IFF

\[ \sum_{i} \frac{p(O|W_i,P_i)}{p(O|W_T,P_T)} \frac{p(W_i,P_i)}{p(W_T,P_T)} < \sum_{i} \frac{p(O|W_i)}{p(O|W_T)} \frac{p(W_i)}{p(W_T)} \]
2. The Normalization Problem

F0, Duration, Energy, Glottal Wave Shape, Spec. Env.:

{Influence of Speaker and Phoneme}

>>

{Influence of Prominence}.

Normalization Problems:
1. Sparse Data
2. MFCCs: influence of phoneme >> prosody
3. F0: influence of speaker ID >> prosody
4. Duration: influence of phoneme >> prosody
Data sparsity

- **Boston Radio News corpus**
  - 7 talkers; Professional radio announcers
  - 24944 words prosodically transcribed
  - Insufficient data to train triphones:
    - Hierarchically clustered states: HERest fails to converge (insufficient data).
    - Fixed number of triphones (3/monophone): WER increases (monophone: 25.1%, triphone: 36.2%)

- **Switchboard**
  - Many talkers; Conversational telephone speech
  - About 1700 words with full prosodic transcription
  - Insufficient to train HMM, but sufficient to test
Proposed solution: Factored models

1. Factored Acoustic Model:
   \[ p(X,Y|Q,A,B) = \prod_i p(d_i|q_i,b_i) \prod_t p(x_t|q_i) p(y_t|q_i,a_i) \]
   - prosody-dependent allophone \( q_i \)
   - pitch accent type \( a_i \in \{\text{Accented}, \text{Unaccented}\} \)
   - intonational phrase position \( b_i \in \{\text{Final}, \text{Nonfinal}\} \)

2. Factored Language Model:
   \[ p(W,P,S) = p(W) p(S|W) p(P|S) \]
Acoustic factor #1: Are the MFCCs Prosody-Dependent?

Clustered Triphones

Prosody-Dependent Allophones

WER: 36.2%

WER: 25.4%

BUT: WER of baseline Monophone system = 25.1%
**Prosody-dependent allophones: ASR clustering matches EPG**

<table>
<thead>
<tr>
<th>Consonant Clusters</th>
<th>Phrase Initial</th>
<th>Phrase Medial</th>
<th>Phrase Final</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accented</td>
<td>Class 1</td>
<td></td>
<td>Class 3</td>
</tr>
<tr>
<td>Unaccented</td>
<td></td>
<td>Class 2</td>
<td></td>
</tr>
</tbody>
</table>

Fougeron & Keating (1997)

EPG Classes:
1. Strengthened
2. Lengthened
3. Neutral
Acoustic Factor #2: Pitch. 
Background: A Speech Synthesis Model of F0 Contours  
(Fujisaki, 2004)
Pitch Pre-Processing for ASR
(Kim, Hasegawa-Johnson, and Chen, 2003)

1. F0 and Probability_Voicing (PV) generated by get_f0 (ESPS)
2. Discard frames with PV < threshold
3. Train an utterance-dependent 3-mixture GMM:
   • \[ p(\log F0 \mid \text{model}) = \sum_{k=-1}^{1} c_k \mathcal{N}\{\log F0; \mu+k\log 2, \sigma^2\} \]
   • Mixture means are \(\mu-\log 2, \mu, \mu+\log 2\)
   • All mixtures have the same variance
4. Discard frames that are \(k=-1\) (pitch halving errors) or \(k=+1\) (pitch doubling errors)
5. Replace missing frames by \textit{linearly interpolating} between good frames
6. \textbf{Log-transform} (to approximate Fujisaki’s F0 model):
   \[ Z_t = [ \log((1+F0_t)/\mu), \log((1+E_t)/\max E_t) ]^T \]
7. \textbf{Discriminative Transform}: \(\chi_t = g(Z_{t-5}, \ldots, Z_t, \ldots, Z_{t+5})\) where \(g()\) is an NN trained to classify frames as pitch accented vs. unaccented.
Acoustic-prosodic observations: 
\[ Y(t) = \text{ANN}(\log f_0(t-5), \ldots, \log f_0(t+5)) \]

Blue Line = Output of the log-transform = Input to the neural network 
Pink Line = Output of the neural network 
Yellow Line = Pitch accent labels used to train the neural network
Acoustic factor #2: Pitch

MFCC Stream

Q(t) = Monophone Labels

Accented? A(t) ∈ {1, 0}

Transformed Pitch Stream
Acoustic Factor #3: Duration

- Normalized phoneme duration is highly correlated with phrase position
- Solution: Semi-Markov model (aka HMM with explicit duration distributions, EDHMM)

\[
P(x(1),...,x(T)|q_1,...,q_N) = \sum_d p(d_1|q_1) \cdots p(d_N|q_N) \\
p(x(1)...x(d_1)|q_1) \cdot p(x(d_1+1)...x(d_1+d_2)|q_2) \cdots
\]
Phrase-final vs. Non-final Durations learned by the EDHMM
A factored language model

Prosodically tagged words:
\[ \text{cats}^{*} \text{ climb trees}^{*\%} \]

1. **Unfactored**: Prosody and word string jointly modeled:
   \[ p( \text{trees}^{*\%} \mid \text{cats}^{*} \text{ climb} ) \]

2. **Factored**:
   - Prosody depends on syntax:
     \[ p( w^{*\%} \mid N V N, w^{*} w ) \]
   - Syntax depends on words:
     \[ p( N V N \mid \text{cats climb trees} ) \]
Result: Syntactic mediation of prosody reduces perplexity and WER

Factored Model:
- Reduces Perplexity by 35%
- Reduces WER by 4%

Syntactic Tags:
- For pitch accent:
  - POS sufficient
- For IP boundary:
  - Parse information useful if available
Syntactic factors: POS, Syntactic phrase boundary depth

![Graph showing prediction errors for accent and boundary with different factors: Chance, POS, POS + Phrase.](image)
Results: Word Error Rate (Radio News Corpus)
Results: Pitch Accent Error Rate

- Chance Recognizer Error Rate
- Radio News, Words Unknown
- Radio News, Words Recognized
- Radio News, Words Known
- Switchboard, Words Known
Results: Intonational Phrase Boundary Error Rate

- Chance
- Radio News, Words Recognized
- Radio News, Words Known
- Switchboard, Words Known

Graph showing the error rates for different categories.
Conclusions

• **Learn from sparse data: factor the model**
  – F0 stream: depends on pitch accent
  – Duration PDF: depends on phrase position
  – POS: predicts pitch accent
  – Syntactic phrase boundary depth: predicts intonational phrase boundaries

• **Word Error Rate**: reduced 12% only if both syntactic and acoustic dependencies modeled

• **Accent Detection Error**:  
  – 17% same corpus words known
  – 21% different corpus or words unknown

• **Boundary Detection Error**:  
  – 7% same corpus words known
  – 15% different corpus or words unknown
Current Work: Switchboard

1. Different statistics ($p_a=0.32$ vs. $p_a=0.55$)
2. Different phenomena (Disfluency)
Current Work: Switchboard

- About 200 short utterances transcribed, and one full conversation. Available at: http://prosody.beckman.uiuc.edu/resources.htm
- Transcribers agree as well or better on Switchboard than on Radio News
  - 95% agreement on whether or not a pitch accent exists
  - 90% agreement on the type of pitch accent (H vs. L)
  - 90% agreement on whether or not a phrase boundary exists
  - 88% agreement on the type of phrase boundary
- Average intonational phrase length is much longer
  - 4-5 words in Radio News
  - 10-12 words in Switchboard
- Intonational Phrases are broken up into many smaller “intermediate phrases:”
  - Intermediate phrase length = 4 words in Radio News; same length in Switchboard
- Fewer words are pitch accented: One per 4 words in Switchboard, vs. one per 2 words in Radio News
- 10% of all words are in the reparandum, edit, or alteration of a DISFLUENCY