Automatic discriminative measurement of voice onset time

Computer Science Graduate Seminar
U. Chicago

Morgan Sonderegger

January 25, 2010
Introduction

▶ Every language’s sound system consists of phones.
    ▶ Stops (/p/, /b/, /q/)
    ▶ Vowels (/a/, /ε/, /i/)
    ▶ Fricatives (/s/, /f/, /ʃ/)
▶ Significant variation in phone realizations, within and across speakers, dialects, languages.
▶ Ex: English {p,t,k} aspirated word-initially (*Pedro, taco*), unaspirated in Spanish.
▶ To study variation, need measurements of perceptually-important quantities over large phonetic corpora.
▶ Scientific interest: How and why does acoustic realization of phones vary?
▶ Practical interest: Variation a huge problem for SR systems.
Manual measurements expensive, impractical for corpora, but automatic measurements must be accurate. (Effect sizes often small.)

Currently, automatic measurement usually not accurate enough:

- ✓: Formants, pitch, speech rate
- ✗: Stress, phone boundaries, nasalization, rhoticity, voice onset time (VOT)

Potential payoff: Manual annotation $\approx$ 3–5 minutes/VOT example, need $> 500$ for a small corpus.

Project: Automatic VOT measurement for initial voiceless stops.
Background

- English stops: /p/, /t/, /k/, /b/, /d/, /g/
- Closure, followed by burst.
- Differ along two dimensions:
  - Voicing
  - Place of articulation

<table>
<thead>
<tr>
<th></th>
<th>Voiced</th>
<th>Voiceless</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lips</td>
<td>/b/</td>
<td>/p/</td>
</tr>
<tr>
<td>Hard palate</td>
<td>/d/</td>
<td>/t/</td>
</tr>
<tr>
<td>Soft palate</td>
<td>/k/</td>
<td>/g/</td>
</tr>
</tbody>
</table>
Background

- **The signal**: Discrete-time sample $y[n]$ of speech signal, $x(t)$.
- **The spectrum**: Usually consider magnitude of short-time (discrete) Fourier transform of $y[n]$ over window centered at $t_1$, $t_2$,...: “power spectrum”.
- (Blackboard interlude)

**Acoustic properties of stops:**
- Closure: Silence
- Burst: Aperiodic, “noisy” spectrum, turbulence
  - Beginning of periodicity = end of stop.

**Humans use many cues to distinguish between stops:**
- Formant transitions to following vowel
- Burst spectrum
- Voice onset time
Voice Onset Time

- VOT = $t$ (onset of voicing) - $t$ (onset of burst).
- Onsets of burst and voicing abrupt, often $\leq 1$ pitch period (5–10 ms).

From T. Neary’s (Alberta) slides, Phonetics
Voice Onset Time

- Important in human (and chinchilla) stop perception

- $p < t < k$, unaspirated $< $ aspirated

- Voiced $< $ unvoiced

- Language, dialect, speaker dependent.

From Kuhl & Miller (1978)
Previous work

Want: Direct comparison of automatic and manual VOT measurements.

   - Data: TIMIT
   - Weapon of choice: Reassignment spectrum

2. Yao (2007):
   - Data: Buckeye
   - WOC: MFCC-based ‘spectral templates’

3. Das & Hansen (2004):
   - Data: Laboratory speech
   - WOC: Teager Energy Operator

- All: Pre-specified rules.

- Current approach: Learned classifiers.
Data

- Word-initial voiceless stops, VOT measured by hand, two sources:
  1. Big Brother corpus: Conversational speech, 4 speakers, British.
  2. TIMIT: Read speech, 630 speakers, American.
- Data: Words with voiceless initial stops, excised from the utterance.
- Goal: Find onsets of burst and voicing ⇒ VOT.
Framework

- **Structured SVM** (Taskar et al. 2003, Tsochantaridis et al. 2004)
  - Generalization of classic SVM (classification, regression) to structured output (sequences).
  - Application to forced alignment of sequences (phonemes, musical notes) in audio: (Keshet et al. 2005, 2007)

- **Data**: Acoustic feature vectors \( \mathbf{X}_i = (\mathbf{x}_1, \ldots, \mathbf{x}_{T_i}) \).
  - \( T_i \): Length of example \( i \) (# of frames).
  - \( d \) features, \( \mathbf{x}_j \in \mathbb{R}^d \).
  - Ex: Energy of power spectrum in window \( j \).

- **Labels**: \( \mathbf{Y}_i = (y_1, y_2) \): Burst and voicing onsets.
  - \( y_i \in 1, \ldots, T_i \).

- **Alignment feature functions**: \( \phi_j(\mathbf{X}, \mathbf{Y}) \)
  - Should be higher for “good” alignments.
  - E.g. total energy in \( (y_1, y_2) \)
Phoneme boundaries are fuzzy ⇒ no perfect alignment.

Cost of predicting $Y = (y_1', y_2')$ for “true” alignment $Y' = (y_1, y_2)$:

$$\gamma(Y, Y') = \sum_{i=1}^{2} \max\{0, |y_i - y_i'| - \epsilon\}$$

Only penalize boundaries more than $\epsilon$ from true boundary.

Today, $\epsilon = 3$.

VOT inter-transcriber agreement usually 0–10 ms.
Goal: Find $w$ s.t.

$$w \cdot (\phi(X_i, Y_i) - \phi(X_i, Y')) \geq \gamma(Y_i, Y')$$

for all $i, Y'$

Intuition: Want one direction in $\phi$ space along which each example's true alignment separated from other alignments by a margin, determined by $\gamma$.

Prediction: Given $w$, alignment for $X$ is

$$f(X; w) = \arg\max_Y w \cdot \phi(X, Y)$$

$f(X; w)$ computation fast, similar to “forward algorithm” for HMMs.

(Some restrictions on $\phi$ apply.)
Finding $w^*$

- Optimization problem:

$$w^* = \arg\min_w \frac{1}{2} \|w\|^2 + C \sum_i \ell(w; (X_i, Y_i))$$

where $\ell$ is a hinge loss:

$$\max\{0, \max_Y [\gamma(Y_i, Y) - w \cdot (\phi(X_i, Y_i) - \phi(X_i, Y))]\}$$

⇒ “violate margin constraints as little as possible”

- Computationally intractable (must consider space of alignments over all $X_i$ jointly)

- Can approximate $w^*$ well and fast by iterative algorithm (Keshet et al. 2007, Crammer et al. 2006):
  - Cycle through training examples $X_i$, find best $w$ for $X_i$ given current estimate ⇒ $w_i$
  - $w^* = \text{best } w_i$ on development set.
Acoustic features \((x_j)\)

- Signal \(y_i[n]\) of word \(i\), beginning with a voiceless stop (closure).
- \(X_i = (x_1, \ldots, x_{T_i})\): \(x_j, x_{j+1}\) separated by 1 ms (frame size).
- Each \(x_j\): 5 features computed over 10 ms window.

1. Energy: \(\log(\text{total E}), \log(\text{high E}), \log(\text{low E})\)
2. Autocorrelation of \(y[n-200 : n+200]\)
   - Max component of power spectrum.
3. Wiener entropy: \(\log(\int E(f)) - \int(\log[E(f)])\)
   - Measures flatness of power spectrum of \(y\).
Example: “talking”
Feature functions ($\phi_j$)

At each possible $Y = (y_1, y_2)$ for $X_i$, 52 feature functions.

Examples:

- Mean of max autocorr component around $y_2$:
  High at voicing onset.

- Mean of Wiener entropy in $(y_1, y_2)$-mean in $(1, y_1)$:
  High at burst onset.

- Mean of log(high E) in $(y_1, y_2)$-mean in $(1, y_1)$:
  High at burst onset.

- “Local difference” $\Delta^F_a(y_i)$ at $y_1, y_2$:
  - Difference between mean of $F$ in $(y_i - a, y_i)$ and $(y_i, y_i + a)$
  - $a \in \{5, 10, 15\}$, $F \in$ all features.

- High when acoustic features change rapidly.

- Approximates derivative, resolution parametrized by $a$. 
Experiments: Preface

- Data: Words beginning with initial voiceless stops (/p/, /t/, /k/).
- $Y_i \in \mathbb{R}^2$: Hand-labeled VOT boundaries.
- Empirically, varying free parameters ($C$, $\epsilon$) changes performance little.
- $\Rightarrow$ fixed $C = 5$, $\epsilon = 2$, no development set.
Experiment 1: TIMIT

TIMIT:
- Read sentences, American speakers, 8 dialect regions.
  - “Birthday parties have cupcakes and ice cream.”
  - “Masquerade parties tax one’s imagination.”
- Stop bursts and closures separately labeled.

Experiment:
- Dialect region 4 only: 657 stops, 100 speakers.
- 80/20 training/test split.
- 5x cross-validation.
Results: TIMIT

|estimated-actual VOT|:

<table>
<thead>
<tr>
<th></th>
<th>≤2 ms</th>
<th>≤5 ms</th>
<th>≤10 ms</th>
<th>≤20 ms</th>
</tr>
</thead>
<tbody>
<tr>
<td>SSVM</td>
<td>23%</td>
<td>53%</td>
<td>78%</td>
<td>97%</td>
</tr>
<tr>
<td>SvH</td>
<td>32%</td>
<td>43%</td>
<td>78%</td>
<td>93%</td>
</tr>
<tr>
<td>FA (baseline)</td>
<td>6%</td>
<td>14%</td>
<td>31%</td>
<td>65%</td>
</tr>
</tbody>
</table>

- **SSVM**: Current approach
- **SvH**: Stouten & van Hamme (2009), *Speech Communication*.
- **FA**: Phone-level forced aligner, standard feature set (MFCCs+Δ+ΔΔ), trained on *all* TIMIT data.
Experiment 2: Big Brother

- Collaborators: Max Bane (U. Chicago), Peter Graff (MIT).
- Conversational speech, British speakers, 4 dialect regions.
  - “I thought I was a party girl till I came in this house, and realized, like, what partying partying partying is.”
  - “I’m gonna miss the camera I’m... going to miss the carpets, I’m even going to miss the cutlery.”
- Two annotators, 85%/15%.

Experiment:
- 704 stops, 4 speakers.
- ≈ 75/25 training/test split.
- Cross validation: Train on 3 speakers, test on fourth.
Results: Big Brother

|estimated-actual VOT|:

<table>
<thead>
<tr>
<th></th>
<th>≤2 ms</th>
<th>≤5 ms</th>
<th>≤10 ms</th>
<th>≤15 ms</th>
<th>≤25 ms</th>
</tr>
</thead>
<tbody>
<tr>
<td>SSVM</td>
<td>49%</td>
<td>77%</td>
<td>88%</td>
<td>92%</td>
<td>97%</td>
</tr>
<tr>
<td>ITR</td>
<td>43%</td>
<td>73%</td>
<td>80%</td>
<td>88%</td>
<td>93%</td>
</tr>
</tbody>
</table>

- Intertranscriber reliability: Comparison of measurements by 2 annotators, on 10% of stops.

⇒ Better than intertranscriber agreement.
Results: Misc

- Other previous work uses different error metrics.
- Attempt at comparison:

<table>
<thead>
<tr>
<th></th>
<th>r</th>
<th>RMS Err. (onset)</th>
<th>Mean Err.</th>
<th>&gt;10% Err.</th>
</tr>
</thead>
<tbody>
<tr>
<td>SSVM (BB)</td>
<td>0.96</td>
<td>7.7 ms</td>
<td>4.0 ms</td>
<td>19.9%</td>
</tr>
<tr>
<td>(TIMIT)</td>
<td>0.93</td>
<td>6.8 ms</td>
<td>6.1 ms</td>
<td>40.0%</td>
</tr>
<tr>
<td>DH</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>25.1%</td>
</tr>
<tr>
<td>YY</td>
<td>—</td>
<td>10.8 ms</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>FA (baseline)</td>
<td>—</td>
<td>—</td>
<td>15.7 ms</td>
<td>77.9%</td>
</tr>
<tr>
<td>Human ITR</td>
<td>0.95–0.99</td>
<td>—</td>
<td>2.9–4.0 ms</td>
<td>—</td>
</tr>
</tbody>
</table>

DH: Das & Hansen (2004), NORSIG.
Humans: Morris et al. (2008), J. Phonetics; Theodore et al. (2009), JASA; Fischer & Goberman (2010), J. Commun Disord.
Discussion

- Exact comparisons not possible, but do well wrt previous work.
- We match (beat?) intertranscriber agreement on BB data.
- TIMIT numbers worse than other datasets, both here and in previous work. (TIMIT transcription problems?)

Todo:
- Voiced stops.
- All positions (not just word initial) ⇒ bursts optional.
- Work from orthographic transcription using forced alignment. (Time consuming to label word boundaries.)
Thanks

- Joseph Keshet
- Karen Livescu
- Max Bane
- Ross Girshick


