Automatic discriminative measurement of voice onset time

Morgan Sonderegger (U. Chicago)

In collaboration with Joseph Keshet (TTI-C)

Illinois Speech Day 2010
Motivation

VOT widely measured in

- **Phonetic research:**
  - **Ex:** Cues to stop voicing/place cross-linguistically (Lisker & Abramson 1964, Cho & Laefoged 1999)

- **Clinical studies:**
  - **Ex:** Diagnosing communication disorders, characterizing disordered speech (Blumstein et al. 1977, 1980; Auzou et al. 2000)

- Currently, measurement (almost) always done **manually:** expensive, hard to scale up.

- However, no sufficiently accurate automatic measurement algorithm.

- **Today:** Automatic VOT measurement for initial voiceless stops.
  - Train algorithm on manual measurements.
  - Evaluate by comparing automatic & manual measurements on new data.
Voice Onset Time

- Important acoustic cue for distinguishing between stops.
- VOT = $t_{\text{onset of voicing}} - t_{\text{onset of burst}}$.
- Onsets of burst and voicing transient, often $\leq 1$ pitch period (5–10 ms).

From T. Neary's (Alberta) course slides
Voice Onset Time

- Perceptually very important →
- p< t< k, unaspirated< aspirated
- voiced< unvoiced

- Varies with speaking rate, stress, word frequency, speaker, language, speaking style...
Previous work on automatic VOT measurement

No automatic/manual comparison:
- ...others?

Include automatic/manual comparison:
- Das & Hansen (2004)
  - Isolated words, using Teager Energy Operator.
- Yao (2007)
  - Buckeye, using MFCC-based ‘spectral templates’.
- Stouten & van Hamme (2009)
  - TIMIT, using time-reassigned spectra.
- All: Customized rules
- Our approach: Discriminative learning algorithm.
Data

- Word-initial voiceless stops.
  1. **Big Brother** corpus: Conversational, 4 speakers, British.
     - VOT measured by hand for 704 stops.
  2. **TIMIT**: Read, 630 speakers, American.
     - Used burst boundaries as VOT.
     - No sa1/sa2
     - Only stops with preceding closure, following voicing (3088/1038 train/test).

- **Data**: Words excised from the utterance
  - Assuming word boundaries known

- **Goal**: Find onsets of burst and voicing $\Rightarrow$ VOT.
Framework

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

- **Data**: Acoustic feature vectors $X_i = (x_1, \ldots, x_{T_i})$.
  - $T_i$: Length of example $i$ (in frames).
  - $x_j \in \mathbb{R}^d$: Energy, entropy, MFCCs (etc.) within a window.

- **Labels**: $Y_i = (y_1, y_2)$: Burst and voicing onsets.
  - $y_i \in 1, \ldots, T_i$.

- **Feature maps**: $\phi_j(X, Y)$
  - Should be higher for “good” VOT.
  - E.g. total energy in $(y_1, y_2)$
Phonetic measurement boundaries are fuzzy ⇒ no perfect label.

- VOT error of prediction $Y = (y_1', y_2')$ for “true” label $Y' = (y_1, y_2)$:
  \[ E_{vot}(Y, Y') = |(y_2' - y_1') - (y_2 - y_1)| \]

- Cost:
  \[ \gamma(Y, Y') = \max\{0, E_{vot}(Y, Y') - \epsilon\} \]

- Only penalize estimated VOT more than $\epsilon$ from “true” VOT.

- Today, $\epsilon = 4$, on the order of VOT inter-transcriber reliability.
Goal: Find \( w \) s.t.

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

for all \( i, Y' \)

Find one direction in \( \phi \) space along which each example's true label separated from other labels by a margin, determined by \( \gamma \).

Prediction: Given \( w \), label for \( x \) is

\[
    f(x; w) = \arg\max_Y w \cdot \phi(x, Y)
\]
Finding $w^*$

- Optimization problem:

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

where $\ell$ is 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”

- Can approximate $w^*$ well and fast by online “passive-aggressive” algorithm (Keshet et al. 2007, c.f. Crammer et al. 2003):
  - Cycle through training examples $X_i$, possibly $M$ epochs.
  - At each step, given $w_i$, find best $w$ for $X_i \Rightarrow w_{i+1}$
  - $w^* = \text{mean } w_i \text{ over training examples.}$
Acoustic features ($x_i$)

- Signal $s[n]$ of a word, beginning with a voiceless stop (closure).
- $X_i = (x_1, \ldots, x_{n_i})$: $x_i$, $x_{i+1}$ separated by 1 ms (frame size).
- Each $x_i$: Features computed over 5 ms window.

1. **Energy**: log(total E), log(high E), log (low E)
2. **Autocorrelation** of $s[n - 100 : n + 200]$
   - Maximum of FFT.
3. **Wiener entropy**: $\log(\int E(f)) - \int (\log[E(f)])$
   - Measures flatness of power spectrum of $y$.
4. **Pitch track, voicing** (0/1)
   - Smoothed with 5 ms Hamming window
Example: “can’t”
Feature maps ($\phi_i$)

For proposed $Y = (y_1, y_2)$ for $X_i$, 59 feature maps.

- Examples:
  - 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.
  - Ex.: Local differences of voicing, autocorrelation at $y_2$:
    High at voicing onset

- Approximates derivative, resolution parametrized by $a$.

... and 51 more.
Experiment 1: Big Brother

- 4 speakers, ≈ 175 VOTs/speaker.
- **Cross-validation**: Train on 3 speakers, test on the 4th.
- \( C = 5, \epsilon = 4, 3 \) epochs.

- To compare to intertranscriber reliability, new transcriber re-measured 100 VOTs.

**Results:**

<table>
<thead>
<tr>
<th>difference in VOTs</th>
<th>( \leq 2 \text{ ms} )</th>
<th>( \leq 5 \text{ ms} )</th>
<th>( \leq 10 \text{ ms} )</th>
<th>( \leq 15 \text{ ms} )</th>
<th>( \leq 25 \text{ ms} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Auto/manual</td>
<td>60.7%</td>
<td>78.4%</td>
<td>87.5%</td>
<td>92.3%</td>
<td>96.0%</td>
</tr>
<tr>
<td>Intertranscriber</td>
<td>54.4%</td>
<td>79.6%</td>
<td>89.3%</td>
<td>93.2%</td>
<td>96.1%</td>
</tr>
</tbody>
</table>

Accuracy close to intertranscriber reliability.
Experiment 2: TIMIT

Train:
  ▶ All initial voiceless stops in TIMIT training set preceded by closure, followed by voiced phone (3088 examples).
  ▶ $C = 5$, $\varepsilon = 4$, 1 epoch.

Test:
  ▶ All such stops in TIMIT core, complete test sets.

Results:

| | automatic–manual VOT |
|---|---|---|---|---|---|
| | $\leq 2$ ms | $\leq 5$ ms | $\leq 10$ ms | $\leq 15$ ms | $\leq 25$ ms |
| Core | 49.5% | 68.8% | 88.2% | 98.9% | 100% |
| Complete | 46.3% | 64.9% | 83.1% | 92.9% | 97.9% |
Comparison with previous work (I)

Stouten & van Hamme (2009), *Speech Communication*

- Manual VOT for 293 voiceless stops from TIMIT.
- *All* stops: includes non-initial, without closure.
- Automatic VOT measurement for each stop: customized rules, using time-reassigned spectra.

- We measure VOT for the same stops, using two weight vectors:
  - $w^*$ trained on all BB data (1 epoch).
  - $w^*$ from Expt. 2 (*TIMIT*)
- Compare automatic & manual agreement:
  - SvH automatic vs. SvH manual
  - TIMIT automatic vs. SvH manual
  - BB automatic vs. SvH manual
- Our algorithm is at a disadvantage: different train/test data, non-initial stops.
Test on SVH data

VOT error (msec)

Cumulative distribution

0.2

0.4

0.6

0.8

1.0

0 10 20 30 40 50

SSVM (TIMIT)

SSVM (BB)

SvH

automatic − manual VOT

<table>
<thead>
<tr>
<th>automatic–manual VOT</th>
</tr>
</thead>
<tbody>
<tr>
<td>≤2 ms</td>
</tr>
<tr>
<td>TIMIT $w^*$</td>
</tr>
<tr>
<td>BB $w^*$</td>
</tr>
<tr>
<td>SvH (2009)</td>
</tr>
</tbody>
</table>
Comparison with previous work (II)

- Das & Hansen (2004): Isolated words
  - Error metric: % VOTs with $\geq 10\%$ error
  - Error metric: RMS err. at onset

Error metrics for Experiments 1–2:

<table>
<thead>
<tr>
<th></th>
<th>RMS (ms)</th>
<th>$\geq 10%$ err.</th>
</tr>
</thead>
<tbody>
<tr>
<td>TIMIT complete</td>
<td>8.66</td>
<td>0.35</td>
</tr>
<tr>
<td>TIMIT core</td>
<td>5.28</td>
<td>0.34</td>
</tr>
<tr>
<td>Big Brother</td>
<td>7.74</td>
<td>0.23</td>
</tr>
<tr>
<td>Yao</td>
<td>10.8</td>
<td>–</td>
</tr>
<tr>
<td>Das/Hansen</td>
<td>–</td>
<td>0.25</td>
</tr>
</tbody>
</table>
Discussion

- Discriminative approach to VOT measurement
- **Learn** weight vector from manually-labeled data.
  - ⇒ allows for different measurement conventions.
- Compares well to previous work, inter-transcriber reliability.

**Future work:**

- **Voiced**, pre-voiced, non-initial stops.
- **Automatic word boundaries** using (adapted) forced alignment.
- Laboratory speech (for phonetic studies).
- Application to VOT in conversational corpus (Big Brother).
Thanks

- Karen Livescu
- Max Bane
- Hugo van Hamme
- Ed King
- Matt Goldrick