Introduction

- **Voice onset time**: Difference between onsets of stop burst and voicing.
- Important perceptual cue to stop voicing & place.
- Clinical, phonetic studies: Hundreds (more?) of transcriber-hours per year.
- Automatic measurement must be:
  - **Trainable**: Differing lab conventions, some subjectivity.
  - **Accurate**: Effect size of interest are small.

Algorithm overview

- **Learn** from manual measurements.
- **Discriminative algorithm**.
  - **Test** task tightly coupled with **training** objective.
  - Trains classifier to minimize the difference $\gamma$ between the manually and automatically-measured VOT.
  - Near-intertranscriber accuracy on conversational speech corpus.
  - Out-performs gold standard (Stouten & van Hamme, 2009).
  - **Code** freely available.

Problem Setting

- **Data**: $\mathbf{x} = (x_1, \ldots, x_T) \in \mathbb{R}^T$.
- **Label**: $(b, t, b_v) \in \mathbb{T}^2$.
- **Goal**: Learn $f : \mathbb{R}^T \to \mathbb{T}^2$.

Learning

- Supervised learning.
- **Cost** function: Predicted VOT within $r$ of manual value not penalized, to allow for variation in "correct" labels.
- **Assume $N$ feature maps** $\phi : \mathbb{R}^T \times \mathbb{T}^2 \to \mathbb{R}$.
- Should be high for small $\gamma$.
- **Learn** $w \in \mathbb{R}^N$:
  - $f(\mathbf{x}) = \arg \max_{(b, t, b_v)} w \cdot \phi(\mathbf{x}, b, t, b_v)$
- Soft SVM: $w = \arg \min_w \frac{1}{2} \|w\|^2 + \sum_{i=1}^m \gamma_i (f(\mathbf{x}_i), t_i, b_v_i)$
- Solved efficiently using **Passive Aggressive algorithm** (Crammer et al., 2006).

Features

- **Data**: $\mathbf{x} = (x_1, \ldots, x_T)$; word beginning with stop.
- Each $x_t^i$: 7 features, taken every 1 ms.
  - **Energy**: $\log(\text{total E})$, $\log(\text{high E})$, $\log(\text{low E})$; 5 ms window.
  - **Autocorrelation**: Max FFT component of ACF, taken over $t = 6$ ms, $t + 18$ ms.
  - Pitch, Voicing: From RAPT pitch tracker.
  - **Wiener entropy**: $\log(f(E(t)) - f(\log(E(t)))$.

Feature maps

- **Local differences** $\Delta^i_j(F)$: Diff. between mean of $F$ in $(t - a, t)$ and $(t, t + a)$.
- **For each possible** $(b, t, b_v)$, 52 feature maps $\phi(\mathbf{x}, b, t, b_v)$.

<table>
<thead>
<tr>
<th>Value at $b$</th>
<th>Value at $t$</th>
<th>Value at $b_v$</th>
<th>Mean, max $\Delta^i_j(F)$</th>
<th>Mean in $(b, t, b_v)$</th>
<th>Mean in $(t - a, t)$</th>
<th>Mean in $(t, t + a)$</th>
<th>Max in $(b, t, b_v)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\log(E_{tot})$</td>
<td>$x$</td>
<td>$x$</td>
<td>$x$</td>
<td>$x$</td>
<td>$x$</td>
<td>$x$</td>
<td>$x$</td>
</tr>
<tr>
<td>$\log(E_{hi})$</td>
<td>$x$</td>
<td>$x$</td>
<td>$x$</td>
<td>$x$</td>
<td>$x$</td>
<td>$x$</td>
<td>$x$</td>
</tr>
<tr>
<td>$\log(E_{lo})$</td>
<td>$x$</td>
<td>$x$</td>
<td>$x$</td>
<td>$x$</td>
<td>$x$</td>
<td>$x$</td>
<td>$x$</td>
</tr>
<tr>
<td>$\Delta_t$</td>
<td>$x$</td>
<td>$x$</td>
<td>$x$</td>
<td>$x$</td>
<td>$x$</td>
<td>$x$</td>
<td>$x$</td>
</tr>
<tr>
<td>$\Delta_{b_v}$</td>
<td>$x$</td>
<td>$x$</td>
<td>$x$</td>
<td>$x$</td>
<td>$x$</td>
<td>$x$</td>
<td>$x$</td>
</tr>
</tbody>
</table>

VOT Background

- **Place of articulation**: Bilabial $\subset$ coronal $\subset$ velar.
- **Native language**: Phonetic properties of L1 influence in bilingual production (Fowler et al., 2008).
- Brazilian Portuguese L1 $\Rightarrow$ shorter VOTs in English.

Data

- **24 Brazilian Portuguese (L1)/English bilinguals**.
  - US residents (min 1 year), fairly high proficiency.
- 24 English native speakers.
- Northwestern undergraduates, no substantial experience with Portuguese, Spanish, Italian.
- 77 target words began with voiceless stops.
- Paired with colored pictures for naming in English.
- Excluded: Tokens not corresponding to intended picture label, disfluent or code-switched.
- 6795 words with initial p/t/k.

Experiment

- **Split data**: train (75%), test (25%).
- **Automatic measurement**: 4-fold cross-validation on train.
- Mixed-effects linear regression models:
  - Fixed: L1 (Portuguese, English), POA (p/t/k).
  - Random: Speaker, Word.
- Fit separate models to auto & manual measurements for train.
- Evaluation: Compare models’ predictions on test data.

Results

- **Regression coefficient**.
- **Comparison** predicted values from the two models on held-out data:
  - **r = 0.99** correlation, 2.4 msec mean absolute difference.
- **As good as intertranscriber agreement**.
  - E.g., $r = 0.978$ in Whiteside et al. (2004).

Conclusion

- **Automatic VOT measurement algorithm** performs similarly to human labelers.
- Experiments here demonstrate the method’s effectiveness on laboratory speech.
- **Code** freely available.

Acknowledgments

This work was supported by a GAANN fellowship to MS, a Northwestern University Graduate Research Grant to NP, and NIH DC007977 and NSF BCS0846147 to MG.