Computing the reliability of acoustic information in speech

Joe Toscano

Beckman Institute for Advanced Science and Technology
University of Illinois at Urbana-Champaign
Acknowledgements

Duane Watson
University of Illinois at Urbana-Champaign

Andrés Buxo-Lugo
University of Illinois at Urbana-Champaign

Bob McMurray
University of Iowa

Funding:
Beckman Institute for Advanced Science and Technology
Speech perception as information processing

*Speech perception*: the process of mapping acoustic information onto meaningful linguistic representations.
Speech perception as information processing

*Speech perception*: the process of mapping acoustic information onto meaningful linguistic representations

Several tools for studying this:

- Eye-tracking & behavioral experiments
- Electrophysiology and neuroimaging
- Computational modeling
Speech perception as information processing

Speech perception

Speech signal

Lexical/semantic information

tart
beach
cat
dart
bus
peach
Speech perception as information processing

Extracting **cues** from the sound signal

**Recognition**

Identifying **categories** from those cues

**Categorization**

**Speech signal**

**Lexical/semantic information**

- tart
- beach
- cat
- dart
- bus
- peach
How reliable are the cues in speech?

What counts as a cue?

- Acoustic features that provide listeners with information about meaningful linguistic differences
- e.g., voice onset time (VOT) as a cue to voicing

[Histogram graph showing number of tokens vs. VOT (ms)]

Allen & Miller (1999), JASA
Speech perception as information processing

Allen & Miller (1999); Beckman et al. (2012); Lisker & Abramson (1964); Image credit: Roke / Wikimedia Commons
Speech perception as information processing

Toscano, McMurray, Dennhardt, & Luck (2010), Psych Sci
Speech perception as information processing

Listeners encode continuous acoustic cues (Toscano et al., 2010)

- **Problem**: Cues do not appear to be invariant in speech
- Example: F1 x F2 vowel space (Peterson & Barney, 1952; Hillenbrand et al., 1995)
Effects of speaking rate on VOT in laboratory speech

Speech perception as information processing
Speech perception as information processing

What types of representations can allow listeners to overcome the problem of lack of invariance?

Several proposed solutions (Ohala, 1996)

- Measure other cues (Stevens & Blumstein, 1978)
- Use higher-order relationships between cues (Sussman et al., 1991)
- Represent speech in terms of articulatory gestures (Fowler, 1984)
- **Use multiple independent cues** (Lisker, 1986)
Combining cues to overcome variability

Idea has been around for a long time: Lisker (1986), 16 cues to word-medial voicing

But it hasn’t been evaluated in detail

Similar proposal in vision (Jacobs, 2002)
  - Multiple cues to 3-dimensional depth/distance
  - e.g., stereopsis vs. atmospheric perspective
  - Successful in explaining observer’s behavior

Can something similar work for speech?
Overview

1) Computing cue reliability
   - Gaussian mixture model (GMM)
   - Cue reliability metric

2) Do listeners weight cues by their reliability?
   - Integration of VOT and vowel length (VL)

3) Can weighting by reliability improve speech recognition?
   - Identifying voicing from multiple cues

4) What factors affect reliability?
   - Participant task engagement
Computing cue reliability

Voice onset time as a cue to voicing

- Clusters corresponding to phonological categories
- Different patterns across languages (Lisker & Abramson, 1964)

Model

- Categories defined by Gaussian distributions
- Mean ($\mu$)
- Standard deviation ($\sigma$)
- Likelihood ($\Phi$)

McMurray, Aslin, & Toscano (2009); Toscano & McMurray (2010)
Computing cue reliability

Dutch

Swedish

English

Thai

Allen & Miller (1999); Beckman et al. (2012); Lisker & Abramson (1964); Image credit: Roke / Wikimedia Commons
Computing cue reliability

Can be trained via unsupervised learning (McMurray, Aslin, & Toscano, 2009)

English VOTs

Spanish VOTs

Thai VOTs

McMurray, Aslin, & Toscano (2009); Toscano & McMurray (2010)
Computing cue reliability

The model can learn the correct categories for a variety of acoustic cues and phonological distinctions across different languages

Makes few assumptions:

- Unsupervised, incremental learning
- Competition between categories
- Small number of parameters (3) used to describe each category

McMurray, Aslin, & Toscano (2009); Toscano & McMurray (2010)
Computing cue reliability

How do we compute reliability in a mixture model?

*Approach used in vision: Weight cues by their reliability* (Jacobs, 2002)
Computing cue reliability

How do we compute reliability in a mixture model?

*Approach used in vision:* Weight cues by their reliability (Jacobs, 2002)
Computing cue reliability

How do we compute reliability in a mixture model?

*Approach used in vision:* Weight cues by their reliability (Jacobs, 2002)
Computing cue reliability

Cue reliability metric that takes into account

- Variance of each category ($\sigma_1 \times \sigma_2$)
- Distance between categories ($\mu_1 - \mu_2$)
- Likelihood of each category ($\phi_1 \times \phi_2$)

Toscano & McMurray (2010); Cog Sci
Computing cue reliability

Cue reliability metric that takes into account

- Variance of each category ($\sigma_1 \times \sigma_2$)
- Distance between categories ($\mu_1 - \mu_2$)
- Likelihood of each category ($\phi_1 \times \phi_2$)

Toscano & McMurray (2010); Cog Sci
Computing cue reliability

Cue reliability metric that takes into account

- Variance of each category ($\sigma_1 \times \sigma_2$)
- Distance between categories ($\mu_1 - \mu_2$)
- Likelihood of each category ($\phi_1 \times \phi_2$)*

$$r_{cue} = \frac{(\phi_1 \times \phi_2)(\mu_1 - \mu_2)}{(\sigma_1 \times \sigma_2)}$$

* Equal likelihood assumed in calculations provided here  

Toscano & McMurray (2010); Cog Sci
Computing cue reliability

Cue reliability metric that takes into account

- Variance of each category ($\sigma_1 \times \sigma_2$)
- Distance between categories ($\mu_1 - \mu_2$)
- Likelihood of each category ($\phi_1 \times \phi_2$)

$$r_{cue} = \frac{(\phi_1 \times \phi_2)(\mu_1 - \mu_2)}{(\sigma_1 \times \sigma_2)}$$

* Equal likelihood assumed in calculations provided here
Overview

1) Computing cue reliability
   - Gaussian mixture model (GMM)
   - Cue reliability metric

2) Do listeners weight cues by their reliability?
   - Integration of VOT and vowel length (VL)

3) Can weighting by reliability improve speech recognition?
   - Identifying voicing from multiple cues

4) What factors affect reliability?
   - Participant task engagement
Weighting by reliability

VOT and vowel length (VL)

Can VL be used as a proxy for speaking rate to overcome variability?

- Do listeners track both cues?
- If so, do they combine them according to their reliability?

Miller, Reeves, & Green (1986); Allen & Miller (1999); Toscano & McMurray (2012)

Reliability of English VOT for word-initial voicing:

3.54

Reliability of vowel length:

0.20
Weighting by reliability

Visual world eye-tracking paradigm

Referent = beach
Competitor = peach
Unrelated = lock; sheep
Weighting by reliability

VOT and vowel length

- Temporally-asynchronous cues
- Effect of each cue over time
- Immediate use of each cue as it becomes available

Toscano & McMurray (2012), *Attn Percep Psychophys*; Toscano & McMurray (submitted)
Weighting by reliability

Cue reliability metric instantiated in Gaussian mixture model (GMM)

- Multiple mixtures corresponding to different acoustic cues
- Maintains continuous information when cues are combined

Reliability of VOT: **3.54**
Reliability of vowel length: **0.20**

Toscano & McMurray (2010), *Cog Sci*; Toscano & McMurray (2012), *Attn Percep Psychophys*
Does the model weight cues similarly to listeners?
Looked at listeners’ categorization responses

Simulations from models trained on VOT and vowel length distributions
- Categorization functions similar to human listeners
- Same size trading relation (within 1 ms VOT)

Toscano & McMurray (2010), Cog Sci; Toscano & McMurray (2012), Attn Percep Psychophys
Weighting by reliability

Similar results for other cues and phonological contrasts

e.g., Japanese fricative/affricate distinction  (Mitani, Kitama, & Sato, 2006)
Weighting by reliability

Eye-tracking experiments show that listeners use both VOT and VL and treat them as independent cues.

Modeling results suggest that cue reliability metric provides a good model for determining how listeners weight cues.

Toscano & McMurray (2010), Cog Sci; Toscano & McMurray (2012), Attn Percep Psychophys
Overview

1) Computing cue reliability
   ▶ Gaussian mixture model (GMM)
   ▶ Cue reliability metric

2) Do listeners weight cues by their reliability?
   ▶ Integration of VOT and vowel length (VL)

3) Can weighting by reliability improve speech recognition?
   ▶ Identifying voicing from multiple cues

4) What factors affect reliability?
   ▶ Participant task engagement
Speech recognition accuracy with multiple cues

Word-medial stop consonant voicing; often cited as having 16 acoustic cues (Lisker, 1986)

How reliable is each cue?

- Acoustic measurements of multiple cues
- Talkers produced all 6 stop consonants (/b,p,d,t,g,k/) in 4 vowel contexts
- Measured 12 potential cues

Toscano & McMurray (in preparation)
Speech recognition accuracy with multiple cues

Results: Cue reliability varied considerably
Speech recognition accuracy with multiple cues

Does combining multiple cues lead to successful recognition?

Examined several models using different combinations of cues

- Invariant-cue model (1)
- Closure cues (3)
- Categorical model (9)
- Lisker (1986) cues (9)
- All measured cues (12)

Trained logistic regression classifiers to identify voicing category of the sounds

Toscano & McMurray (in preparation)
Speech recognition accuracy with multiple cues

Model performance by vowel

Performance by talker
Speech recognition accuracy with multiple cues

Similar performance for other distinctions

Fricative place of articulation (McMurray & Jongman, 2010)
Speech recognition accuracy with multiple cues

Cues vary in their reliability

Combining multiple cues can overcome variability inherent in individual ones

Cue-integration approach produces results close to human performance

Toscano & McMurray (in preparation)
Overview

1) Computing cue reliability
   - Gaussian mixture model (GMM)
   - Cue reliability metric

2) Do listeners weight cues by their reliability?
   - Integration of VOT and vowel length (VL)

3) Can weighting by reliability improve speech recognition?
   - Identifying voicing from multiple cues

4) What factors affect reliability?
   - Participant task engagement
Task effects on cue reliability

Acoustic cues to prosodic prominence (e.g., *new* vs. *given* information)

Individual cues (e.g., mean F0) seem to provide some information, but are often unreliable in lab experiments

Are these cues really as unreliable as they seem to be?

Prototypical lab task

**Given:**
- Sequence 1: *white, brown, green*
- Sequence 2: *white, brown, green*

**Contrastive:**
- Sequence 1: *white, red, green*
- Sequence 2: *white, brown, green*

**New:**
- Sequence 1: *blue, grey, black*
- Sequence 2: *white, brown, green*

Toscano, Buxo-Lugo, & Watson (forthcoming chapter); Buxo-Lugo, Toscano, & Watson (in prep)
Task effects on cue reliability

Acoustic cues to prosodic prominence (e.g., *new* vs. *given* information)

Individual cues (e.g., mean F0) seem to provide some information, but are often unreliable in lab experiments

Are these cues really as unreliable as they seem to be?

Prototypical lab task

Given:

Sequence 1: *white, brown, green*

**Sequence 2: white, brown, green**

Contrastive:

Sequence 1: *white, red, green*

**Sequence 2: white, brown, green**

New:

Sequence 1: *blue, grey, black*

**Sequence 2: white, brown, green**

Toscano, Buxo-Lugo, & Watson (forthcoming chapter); Buxo-Lugo, Toscano, & Watson (in prep)
Task effects on cue reliability

Acoustic cues to prosodic prominence (e.g., new vs. given information)

Individual cues (e.g., mean F0) seem to provide some information, but are often unreliable in lab experiments

Are these cues really as unreliable as they seem to be?

Prototypical lab task

Computer-game task

Toscano, Buxo-Lugo, & Watson (forthcoming chapter); Buxo-Lugo, Toscano, & Watson (in prep)
Task effects on cue reliability

Measured four cues: (1) Duration, (2) Intensity, (3) Mean F0, (4) F0 range

Toscano, Buxo-Lugo, & Watson (forthcoming chapter); Buxo-Lugo, Toscano, & Watson (in prep)
Task effects on cue reliability

Computed overall cue reliability for each task

\[ r_{\text{cue}} = \frac{(\mu_1 - \mu_2)}{(\sigma_1 \cdot \sigma_2)} \]

Compared with chance reliability estimated via Monte Carlo simulations

- Prototypical lab task: No different from chance (p>0.1)
- Computer-game task: Significantly more reliable (p<0.001)

<table>
<thead>
<tr>
<th>Task</th>
<th>Log-duration</th>
<th>Intensity</th>
<th>Mean F0</th>
<th>F0 Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prototypical task</td>
<td>0.56</td>
<td>0.26</td>
<td>0.31</td>
<td>0.48</td>
</tr>
<tr>
<td>Computer game</td>
<td>0.25</td>
<td>0.15</td>
<td>0.12</td>
<td>0.03</td>
</tr>
</tbody>
</table>

Toscano, Buxo-Lugo, & Watson (forthcoming chapter); Buxo-Lugo, Toscano, & Watson (in prep)
Task effects on cue reliability

Level of engagement has an effect on cue reliability

Shows that effects of lab speech can go in both directions

- Laboratory speech can artificially increase cue reliability (e.g., VOT)
- It can also artificially decrease cue reliability (prosodic prominence)

Toscano, Buxo-Lugo, & Watson (forthcoming chapter); Buxo-Lugo, Toscano, & Watson (in prep)
Thinking about speech perception as an information processing problem that involves combining multiple sources of information:

- Modeling work provides a method for computing cue reliability.
- Listeners' responses suggest that they weight cues according to reliability.
- Classifier demonstrates that integrating multiple cues can lead to successful speech recognition.
- Allows us to study factors affecting reliability (e.g., laboratory speech vs. naturalistic language).

**Combining cues based on their reliability provides a way for listeners to overcome variability in speech**