Categorical Perception and Speech Sound Types

Tsan Huang
thuang3@buffalo.edu

Department of Linguistics
University at Buffalo (SUNY)
Huang & Johnson (2010) – procedure

- Natural speech Mandarin tones T55, T35, T214 and T51
  - monosyllables
  - single talker

- tasks
  - Speeded AX discrimination
  - AX difference rating

- ISI = 100ms in both tasks

- Native Mandarin– and American English (AE)–speaking listeners
Huang & Johnson (2010) – AX difference rating results

Y-axis: The higher the rating, the more different the tone pair.

AE: (lighter gray)
- two good pairs
  (T55/T214; T214/T51; i.e., a H vs. L contrast)
- contour not important
  (see T55/T51, T35/T51)

Chinese:
- two weak contrasts
  (both involved in sandhi)
- contour very important
  (see T55/T51, T35/T51)
– Task easy for everyone, no difference in percentage correct

– The sandhi pair harder for the Chinese

– Auditory vs. linguistic processing: not one or the other, but more one than the other, in a particular task

– For native listeners, tone categories are well defined, except for that between the sandhi pair
Paper title:

“The end of categorical perception as we know it”
Liberman, A., K. Harris, H. Hoffman, & B. Griffith (1957) – stimuli
Identification (labeling) of each stimulus presented in isolation

- Discrimination of stimuli presented in one-, two- and three-step pairs
Hypothesis: When presented with stimuli from a continuum, the listener categorizes them into existing categories in his language.

The extreme assumption (p.362): “S can discriminate the stimuli only to the extent that he can identify them as different phonemes.”
- Discrimination function should be completely predictable from identification function
Liberman, A., K. Harris, H. Hoffman, & B. Griffith (1957) – results
The predicted curves underestimated the listeners’ real discrimination function.

“One possibility … is that this discrepancy represents … an ability of S to distinguish the speech sounds, not solely on the basis of the phonemic labels, but also more directly by the essential acoustic differences among the patterns.” (p.366)
Fry, Abramson, Eimas & Liberman (1962) – procedure

- Thirteen (13) synthetic (two-formant) vowel continuum of {ɪ–ɛ–æ}

- ABX discrimination

- Identification (triads; label each stimulus as one of the three vowels /ɪ; ɛ; æ/)

- Eight (8) American English speakers/listeners
Fry, Abramson, Eimas & Liberman (1962) – identification results
Listeners outperformed the predicted discrimination function by a large margin. Bad stimuli.
American English & Swedish listeners

two synthetic (three-formant) vowel continua of thirteen (13) stimuli each ranging through
- AE /i–ɪ–ɛ/ (or Swedish /i–e–ä/)
- Swedish /i–y–ʉ/.

Identification

ABX discrimination
Stevens, Liberman, Studdert-Kennedy & Ohman (1969) - discrimination results
(p.21) “For the stops there appears to be a relatively complex relation between the phoneme and its representation as sound, and the acoustic cues reside in a signal characterized by rapid changes.”
   ◦ [But plosive voicing cues are not.]

The acoustic cues for the steady–state vowels are relatively simple.

Perhaps perception of “vowels rapidly articulated in normal phonetic contexts” would show a more categorical pattern?
Chan, Chuan & Wang (1975) and Wang (1976): \{i35 – i55\}
Chan, Chuan & Wang (1975) and Wang (1976)

A. Chinese listeners
- Good identification boundary
- discrimination peak aligned with identification boundary

B. AE listeners
- later and less sharp identification boundary
- discrimination peaked even later at the last stimulus pair

Native tone perception appears to be categorical.
Sixteen (16) F0 levels ranging from 92Hz to 152Hz in steps of 4Hz superimposed on the syllable /kha:/

Identification
- Identify the 16 stimulus tones as one of five possible words
- Real targets: three words with low, mid and high level tones

Discrimination:
- 4–interval forced choice (4IAX)
Sun & Huang (under review) – procedures

- stimulus continua synthesized from natural speech tones
  - /si55/ as the starting point in both continua
  - duration and intensity of /si33/ and /si51/ normalized to those of /si55/.

The Ilse Lehiste Memorial Symposium, Nov. 11–12, 2011; Columbus, Ohio
Sun & Huang (under review) – results: \textbf{T55–T51 discrimination (d’)}

\textbf{Two–step discrimination}

\begin{figure}
\centering
\includegraphics[width=\textwidth]{two_step.png}
\end{figure}

\textbf{Three–step}

\begin{figure}
\centering
\includegraphics[width=\textwidth]{three_step.png}
\end{figure}

\textbf{Four–step}

\begin{figure}
\centering
\includegraphics[width=\textwidth]{four_step.png}
\end{figure}

\textbf{Native}:
- Discrimination peak in two–step
- Less prominent peak in three–step
- No peak in four–step

\textbf{AE}:
- No peak

The Ilse Lehiste Memorial Symposium, Nov. 11–12, 2011; Columbus, Ohio
The Ilse Lehiste Memorial Symposium, Nov. 11–12, 2011; Columbus, Ohio

Sun & Huang (under review) – results: T55–T33 discrimination (d’)

Two-step discrimination

Three-step

Four-step

No peak at all in T55–T33 data.

Contour tone perception is somewhat categorical, while level tones show a continuous pattern.

Wang et al. and Abramson were both right.
Gerrits & Schouten (1998) – stimuli

- Vowels presented in /p–V–p/ context
- End points of vowel continuum: [u] and [i]
- “over-articulated vowel continuum” generated with vowels produced in isolated words
- “underspecified vowel continuum” created from vowels produced in a fast read text
“Surprisingly”, no difference between two continua temporal and spectral reduction in “underspecified” vowels did not lead to “more complex coding” or “more categorical perception”.

Gerrits & Schouten (1998) – results
Peterson & Barney (1952): F1–F2 plot of 76 men, women and children

Overlaps:
some /æ/ in /ɛ/ eclipse

/ɜ/ eclipse overlaps with those of /ɛ, ae, ʊ/;

/ɑ/ eclipse intersects those of /ʌ, ɔ/; …

In general, tokens scatter evenly over a large area within each category.

With temporal and spectral reduction, things will be even more crowded.
Similar overlap & within category scatter in tone distribution
Huang (unpublished production study)

same H tone, different speakers

same L tone, different speakers

The Ilse Lehiste Memorial Symposium,
Nov. 11–12, 2011; Columbus, Ohio
Keating (1984): stop consonant voicing (see also Lisker & Abramson, 1964)

Polish /b/ versus /p/: two separate VOT distribution populations
## Inter-talker Variations in Cs, Vs and Tones

<table>
<thead>
<tr>
<th>V formant / tone</th>
<th>VOT</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Within-category scatter</strong></td>
<td>Normal distributions with well-defined peaks</td>
</tr>
<tr>
<td><strong>Between-category overlaps</strong></td>
<td>Clear category boundaries (coinciding with regions of higher auditory sensitivity?)</td>
</tr>
<tr>
<td><strong>Dependent on talker physiology</strong></td>
<td>Relatively independent of ...</td>
</tr>
</tbody>
</table>

- Dialectal differences and speaking habits may contribute more to V formant or T pitch scatter.

- As a result, adequate stimulus–extrinsic information will be necessary for correct interpretation of intrinsic cues in V and T stimuli.
Ladefoged & Broadbent (1957): effect of auditory context on vowel perception

- Depending on the precursor phrase, a stimulus was heard 87% of the time as ‘bit’ or 90% of the time as ‘bet’
Lehiste & Meltzer (1973) – mismatched $f_0$ and formants

Could it be speaker identity?

- Lower vowel perception accuracy when male vowel formants were combined with children’s high pitch & vice versa
Carrier phrase $f_0$ as a cue for speaker identity and a basis for normalization in V perception
- More so than intrinsic $f_0$ in test word
Tone identification boundary shifted with change in $f_0$ range in precursor phrase

More accurate tone identification when stimuli were blocked by talker than when mixed
Given the distribution pattern of V formants and Ts, the listener has to be sensitive to any value within the possible ranges, as each value may represent a particular speaker’s V/T category.

- Evidence from L2 acquisition of English Vs by L1 German and Spanish learners (Iverson & Evans, 2007)

There must be multiple exemplars for each category – or, a flexible category boundary, adjusted every time the listener encounters a new speaker with a different set of V formant or T pitch values.
As a result, the categorical perception test with stimuli drawn from a continuum will not be the right one for V or T perception.

- It will be the end of it in V or T perception!
Contour vs. level tones

But if one does use the paradigm ...

- Since pitch **contour**, which is to some extent less dependent on talker physiology than **pitch height**, one may predict that *caeteris paribus*, contour Ts will be perceived in a more categorical fashion (*Sun & Huang, under review)*.
Monophthong and diphthong Vs

- By the same token, *caeteris peribus*, diphthongs with intrinsic formant ‘contours’ will be perceived in a less continuous pattern than monophthongs.

- But both T and V perception will be more continuous than that of C place or voicing contrast.