# Optimality-theoretic modelling of acoustic cue integration

Paul Boersma (University of Amsterdam) and Paola Escudero (McGill University)

## HYPOTHESES

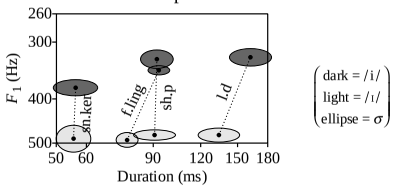
1. The "same" phonological contrast is **produced differently** in different dialects.
2. The "same" phonological contrast is **perceived differently** in different dialects.
3. The perception difference depends on the production difference, i.e., listeners use an **optimal perception strategy** (maximum likelihood), which minimizes the amount of perceptual confusion.
4. We can model the knowledge behind optimal perception and its acquisition with Optimality Theory (OT) and the Gradual Learning Algorithm (GLA).

We test these hypotheses for the English vowel categories /ɪ/ and /i/, which contrast acoustically in **first formant** (F1, "height") and **duration**. The two dialects considered are the standard variants of **Scottish English** and **Southern British English**.

## ATTESTED ADULT PRODUCTION OF /ɪ/ VERSUS /i/

**Experiment:** a male speaker from each dialect produced 50 tokens of each of the eight words *ship, sheep, lid, lead, Snicker, sneaker, filling,* and *feeling* in a carrier sentence.

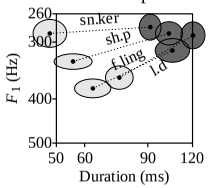
**Results:** **Scottish** speaker







**Southern** speaker

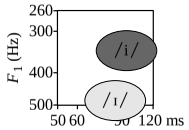



**Observation:** for contrasting /ɪ/ from /i/, Scottish speakers use almost exclusively the spectral (F1) cue, Southern speakers use the spectral as well as the duration cue.

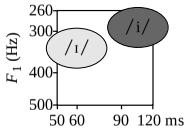
## OPTIMAL PERCEPTION: MAXIMUM LIKELIHOOD

To minimize the probability of misunderstanding the speaker, an optimal listener should perceive any incoming acoustic event as the category that was most likely to have been intended by the speaker. Consider the following production distributions:

**Scottish** environment



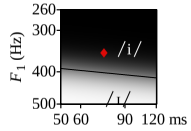

**Southern** environment




(the **means** are based on averaging over all four English word pairs;
the **standard deviations** are informed guesses: $\sigma_{F1}$ = 0.2 octave; $\sigma_{dur}$ = 0.4 doubling)

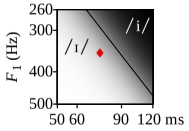
The most likely vowel, given an [F1, duration] pair, then depends on the dialect:

**Scottish** most likely vowel





**Southern** most likely vowel




**Observation:** Scottish optimal listeners should rely almost exclusively on F1, Southern optimal listeners should rely on both duration and F1.

**Example:** the acoustic event "♦", i.e. [349 Hz, 74 ms], should be perceived:
as /i/ by an optimal Scottish listener,
as /ɪ/ by an optimal Southern English listener.

## MODELLING PERCEPTION

The listener's implicit knowledge behind the integration of the two acoustic cues is modelled as an Optimality-Theoretic **perception grammar**, which maps raw acoustic events on discrete arbitrary symbols. The constraints handle the two cues separately:

"an F1 of 260 Hz should not be perceived as /ɪ/"
"an F1 of 260 Hz should not be perceived as /i/"
"an F1 of 500 Hz should not be perceived as /ɪ/"
"an F1 of 500 Hz should not be perceived as /i/"
"a duration of 50 ms should not be perceived as /ɪ/"
"a duration of 50 ms should not be perceived as /i/"
"a duration of 120 ms should not be perceived as /ɪ/"
"a duration of 120 ms should not be perceived as /i/"

Analogous constraints exist for all other values of F1 and duration.

The acoustic event "♦" [349 Hz, 74 ms] is perceived differently in the two dialects, because they have different constraint rankings (only relevant constraints shown):

**Scottish listeners perceive ♦ as /i/:**

| [349 Hz, 74 ms] | | 349 Hz is not /ɪ/ | 74 ms is not /i/ | 74 ms is not /ɪ/ | 349 Hz is not /i/ |
|---|---|---|---|---|---|
| | /ɪ/ | *! | | * | |
| ☞ | /i/ | | * | | * |

**Southern English listeners perceive ♦ as /ɪ/:**

| [349 Hz, 74 ms] | | 349 Hz is not /i/ | 74 ms is not /i/ | 74 ms is not /ɪ/ | 349 Hz is not /ɪ/ |
|---|---|---|---|---|---|
| ☞ | /ɪ/ | | | * | * |
| | /i/ | *! | * | | |

## MODELLING THE PERCEPTUAL ACQUISITION PROCESS

The ranking of the constraints is learned by the Gradual Learning Algorithm. Suppose that a Scottish learner, at some point during acquisition, has a ranking that would be appropriate for an adult Southerner instead. When confronted with the acoustic event "♦" [349 Hz, 74 ms], she will make a mistake:

| [349 Hz, 74 ms] | | 349 Hz is not /i/ | 74 ms is not /i/ | 74 ms is not /ɪ/ | 349 Hz is not /ɪ/ |
|---|---|---|---|---|---|
| *☞* | /ɪ/ | | | ←* | ←* |
| √ | /i/ | *!→ | *→ | | |

If the learner notices her mistake (because she perceived /ʃɪp/ but notices that the semantic context requires the recognition of |ʃip| 'sheep'), she will take action:
- by lowering the ranking of all constraints violated in the correct adult form (√),
- and raising the ranking of all constraints violated in her own incorrect form (*☞*),

by a small amount along the continuous ranking scale.

### Further reading

Escudero, P. & Boersma, P. (to appear): "Modelling the perceptual development of phonological contrasts with Optimality Theory and the Gradual Learning Algorithm." In *Proceedings of the 25th Penn Linguistics Colloquium.* [Rutgers Optimality Archive ROA-439, http://roa.rutgers.edu]

All production and perception data and simulation scripts are available at:
http://www.fon.hum.uva.nl/paul/p2/

The simulations were performed with the *Praat* program, www.praat.org.

Part of this work was done at the Universities of Edinburgh and Reading.

## CONCLUSIONS

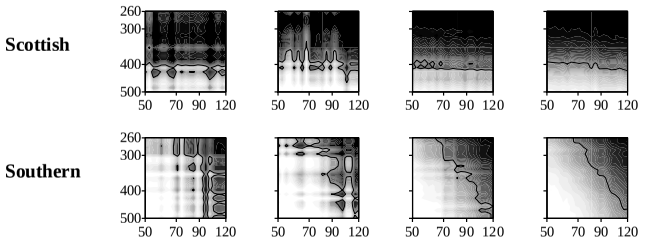
1. Production and perception of the /ɪ/-/i/ contrast are dialect-specific.
2. Real listeners implement an optimal strategy for perception.
3. OT & GLA successfully model optimal perception and its acquisition.

Further modelling and studies:
- Category emergence/split/merger.
- Integration of multiple cues to multiple contrasts.
- Longitudinal and second-language data.

## SIMULATION OF PERCEPTUAL ACQUISITION OF /ɪ/ VS. /i/

In order to see whether our constraint set can actually handle an optimal integration of the two acoustic cues, we created two **virtual listeners** (one Scottish, one Southern) who grow up in the "average" language environments defined in the first column of this poster. In the initial state of these learners, all constraints were ranked at the same height, i.e. our virtual baby learners would perceive every acoustic event as /ɪ/ or /i/ with 50% probability. We then repeatedly fed our virtual listeners with input-output pairs drawn randomly from their respective environments. At several virtual times, we measured the performance of their perception grammars (black is /i/, white is /ɪ/):

**Results:** after... 1 month 10 months 100 months 1000 months

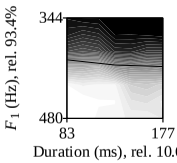



**Observation:** cue reliances for the final stages compare well with the optimal ones. Our model indeed leads to an optimal (maximum-likelihood-like) listener.

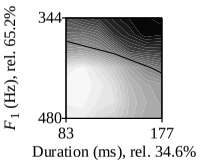
## ATTESTED ADULT PERCEPTION OF /ɪ/ VERSUS /i/

**Experiment:** 20 Scottish and 21 Southern listeners had to classify 370 acoustic events with an F1 between 344 and 480 Hz and a duration between 83 and 177 ms as /ɪ/ or /i/.

**Results:** **Scottish** listeners




**Southern** listeners




**Observation:** real listeners act qualitatively like the optimal and the simulated listeners. Quantitative differences between simulation and experiment may have various causes:

- For the Southerners, the contrast was spectrally enhanced in an unnatural way;
- The spectral cue became available to the listener before the duration cue;
- The stimuli had to be isolated vowels rather than natural speech;
- The simulations are sensitive to the guessed standard deviations ($\sigma_{F1}$ and $\sigma_{dur}$);
- Real listeners have contact with other dialects than the one they speak;
- The Southern speaker de-stressed the target words better than the Scottish speaker;
- It is unknown how representative the two speakers were of their dialects.