One of the tasks of the language acquisition process is to optimize strategies for comprehension. For speech perception, this means that the learner has to establish an accurate mapping from acoustic categories to perceptual categories. As an example, this paper considers the development of the perception of the English vowels /I/ and /i/ in native speakers. Production-wise, the two vowels differ in various respects. In this paper, we will limit ourselves to considering duration and F1 (first formant). It turns out (§1) that the use of these two acoustic dimensions in production depends on the dialect at hand: for Scottish English speakers, /I/ and /i/ differ much more in F1 and much less in duration than for Southern English speakers. We find, therefore, that the perceptual strategies for contrasting /I/ and /i/ in native speakers are different. In this paper, we hypothesize that humans have an optimal strategy for acoustic cue integration. Perception experiments show that the Scot comes to rely almost exclusively on height (F1) when and show that the Southern comes to rely almost exclusively on height (F1) and show that the Scot comes to rely almost exclusively on height (F1) and show that the Southern comes to rely almost exclusively on height (F1) and show that the Scot comes to rely almost exclusively on height (F1).
1.1 The Production Experiment

We recorded fifty tokens of each of the words ship, sheep, filling, feeling, Snicker, sneaker, lid, and lead in the carrier sentence THIS is a __ as well, spoken by a male speaker of Scottish English and a male speaker of Southern English. These words were chosen in order to respect the voicing of the following consonant and the number of syllables. We told the speakers to stress the word THIS, expecting them to destress the target words. There were also ten distractor words, which were recorded ten times each: car, bicycle, chair, kitchen, pad, lip, speaker, mailing, warning, and table.

In total, each speaker pronounced 500 sentences, in about 30 minutes. The words were put in a semi-random order, with all eight target words occurring in every decade. For example, the first ten words were lead, Snicker, ship, feeling, car, sneaker, lid, sheep, filling, and bicycle; the next decade would have the target words in a different order, but the members of each pair were always separated by a distractor word. The speaker would sit at a table with a microphone, and the carrier phrase was stuck to this table. The words were written on 500 cards. The speaker was first asked to say two sets of ten words, and the experimenter wrote down any hesitations. If the speaker hesitated at any words, these words were recorded again afterwards.

1.2 Results of the Production Experiment

The vowels were segmented by both of us separately with the help of the Praat program. The averages of our time markings were used for an automatic analysis of duration and first formant. The results are in Tables 1 and 2. The target words are expressed in base-2 logarithmic units.

A first difference between the two dialects is found in the way the two acoustic dimensions correlate with other factors than the vowel contrast. We observe (Fig. 1) that for the Scottish English speaker, the vowel variation is the primary factor for the voiced consonant, whereas it is the primary factor for the initial consonant of the following consonant.

\[
\begin{array}{c|c|c}
\text{Word} & \text{Duration (ms)} & \text{F1 (Hz)} \\
\hline
\text{ship} & 90.7 & 480 \\
\text{sheep} & 92.0 & 327 \\
\text{filling} & 76.8 & 492 \\
\text{feeling} & 93.1 & 346 \\
\text{Snicker} & 55.5 & 489 \\
\text{sneaker} & 56.2 & 378 \\
\text{lid} & 134.0 & 480 \\
\text{lead} & 162.2 & 324 \\
\end{array}
\]

Fig. 1: Scottish (left) and Southern (right) production of /I/ (light) and /i/ (dark). The axes are logarithmic. The ellipses show the standard deviations.

1.3 Relative Cue Use

Our modelling of the perception of the /I/-/i/ distinction (§2) will be based on the availability of duration and F1 cues in the different production environments. Therefore, we have to accurately compare the Scottish and the Southern speaker with respect to their relative use of the two acoustic dimensions.
In order to single out the correlation between the vowel contrast and the two acoustic dimensions, we average (geometrically) the duration and F1 values for the two vowels in the two dialects ... across number of syllables (one or two). The averaged data are shown in Table 3 and Figure 2.

We can now propose a numeric characterization of a speaker's relative use of the two acoustic dimensions. It is expressed in terms of the horizontal and vertical distances between the two vowels in the duration-F1 plane. In going from Scottish \[ /I/ \] to \[ /i/ \], the mean F1 falls from 485 to 343 Hz, which is 0.500 octaves, while the mean duration rises from 84.8 to 94.0 ms, which is 0.149 duration doublings. This can lead us to define a spectral/duration cue-use ratio of \[ -0.500/0.149 = -3.4 \text{ oct/dur doubling} \]. This number is equal to the slope of the imaginary line that connects the Scottish \[ /I/ \] and \[ /i/ \] in Figure 2. For the Southerner, F1 falls by 0.207 octaves, while the duration rises by 0.809 doublings, so that his cue-use ratio is \[ -0.26 \text{ oct/dur doubling} \]. Apparently, the Scot prefers the F1 dimension (or disfavours the duration dimension) 13 times more than the Southerner does.

In order to model the perception process and its acquisition, we assume that the vowels in the experiment and listener input are drawn from Gaussian distributions centered about the mean F1 and duration values for the Scottish and Southern English vowels. This section presents our model of perceptual development, illustrated by the behavior of virtual Elspeth and virtual Liz, who grow up in virtual Scottish and Southern English environments. In §2.1, the virtual production environment is described. In the next section, we show how the virtual speakers perceive the speech they hear. We then show how the model predicts the vowels in the productions of these virtual speakers. Finally, in §2.3, we present a computer simulation of our model and show that the predictions are borne out by the behavior of real listeners.

Table 3: Duration and F1 for \[ /I/ \] and \[ /i/ \] for the Scottish and Southern English speakers, averaged across the four contexts and the total standard deviations.

<table>
<thead>
<tr>
<th>Context</th>
<th>Scottish</th>
<th>Southern</th>
</tr>
</thead>
<tbody>
<tr>
<td>F1 (Hz)</td>
<td>Duration (ms)</td>
<td></td>
</tr>
<tr>
<td>/I/</td>
<td>485</td>
<td>84.8</td>
</tr>
<tr>
<td>/i/</td>
<td>343</td>
<td>94.0</td>
</tr>
<tr>
<td>Error</td>
<td>0.066</td>
<td>0.058</td>
</tr>
</tbody>
</table>

Figure 2: Duration and F1 for \[ /I/ \] and \[ /i/ \] for the Scottish and Southern English speaker, averaged across the four contexts.
by making decisions that lead to maximum-likelihood behaviour in perception. For speech perception, this means that the best thing for the listener to do is to perceive any incoming acoustic event as the phonological category that was most likely to have been intended by the speaker. We hypothesize that listeners minimize the probability of miscomprehension.

2.2 The Optimal Perception Process

As proposed by Boersma (1998:164), we can account for the mapping of acoustic cues to phonological categories using a formal grammar. This perception grammar contains constraints for mapping acoustic cues to phonological categories. In the case at hand, we label the categories arbitrarily as /I/ and /i/—phonological categories. In the case at hand, we label the categories arbitrarily as /I/ and /i/. We could now test our hypothesis against real listening experiments.

We hypothesize that listeners minimize the probability of miscomprehension.

So how do Elspeth and Liz implement an optimal perception strategy? Our answer is that the knowledge behind their perception process is a formal grammar. This perception grammar contains constraints for mapping acoustic cues to phonological categories. In the case at hand, we label the categories arbitrarily as /I/ and /i/.

2.3 Modelling the Perception Process

However, we believe that there is need to explain in detail the knowledge that underlies overt perceptual behaviour. We will present a model that answers the question: how do listeners implement an optimal perception strategy, and how do they do it? We need to explain the process of decision making in the perception environment, as well as the way in which the model emerges from the interaction between the production environment and the perception environment.

We define the optimal perceiver as the listener to do it, who will rely mainly on reliable cues, will rely almost exclusively on F1 and hardly on duration, will rely on duration primarily, on F1 secondarily. We will later learn how to do it in detail. We need to explain the process of decision making in the production environment.

Figure 3: The production environments for virtual Scottish Elspeth (left) and Southern English Liz (right).

Now that we have values for the standard deviations, we can establish a numeric measure for the reliability of the two cues that signal the same event as /I/ or as /i/.
knowledge of probabilities. Her only knowledge resides in the rankings of the constraints, and any apparent optimal behaviour is derived from that. We should note that in stochastic OT, the listener has no direct knowledge of the constraints. Boersma (1998:338) nevertheless showed that in the case of single-cue categorization, this algorithm leads to a desirable near-optimal property of the algorithm.

Tableau 3: Error-driven learning by the Gradual Learning Algorithm in a stochastic OT environment.

<table>
<thead>
<tr>
<th>Constraints</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>[349 Hz, 74 ms]</td>
<td>1</td>
</tr>
<tr>
<td>[349 Hz, 74 ms]</td>
<td>1</td>
</tr>
<tr>
<td>[349 Hz, 74 ms]</td>
<td>1</td>
</tr>
<tr>
<td>[349 Hz, 74 ms]</td>
<td>1</td>
</tr>
<tr>
<td>[349 Hz, 74 ms]</td>
<td>1</td>
</tr>
<tr>
<td>[349 Hz, 74 ms]</td>
<td>1</td>
</tr>
<tr>
<td>[349 Hz, 74 ms]</td>
<td>1</td>
</tr>
<tr>
<td>[349 Hz, 74 ms]</td>
<td>1</td>
</tr>
</tbody>
</table>

The knowledge underlying the perception of the same acoustic event for Liz, who lives in a Southern English production environment.

Tableau 2: The perception of the acoustic event [349 Hz, 74 ms] for Liz, who is shown in Tableau 2. Her two F1 constraints are ranked in the reverse order from Elspeth's, and she will choose to perceive 'sheep' as the winner (i.e. as the actually perceived category) because this candidate violates the least high-ranked constraints.

Tableau 1: The perception of the acoustic event [349 Hz, 74 ms] for Elspeth, who lives in a Scottish English production environment.

The knowledge underlying the perception of the same acoustic event for Elspeth.

Tableau 3: Error-driven learning by the Gradual Learning Algorithm in a stochastic OT environment.

<table>
<thead>
<tr>
<th>Constraints</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>[349 Hz, 74 ms]</td>
<td>1</td>
</tr>
<tr>
<td>[349 Hz, 74 ms]</td>
<td>1</td>
</tr>
<tr>
<td>[349 Hz, 74 ms]</td>
<td>1</td>
</tr>
<tr>
<td>[349 Hz, 74 ms]</td>
<td>1</td>
</tr>
<tr>
<td>[349 Hz, 74 ms]</td>
<td>1</td>
</tr>
<tr>
<td>[349 Hz, 74 ms]</td>
<td>1</td>
</tr>
<tr>
<td>[349 Hz, 74 ms]</td>
<td>1</td>
</tr>
<tr>
<td>[349 Hz, 74 ms]</td>
<td>1</td>
</tr>
</tbody>
</table>

The knowledge underlying the perception of the same acoustic event for Elspeth.
We simulated the development of a Scottish and a Southern English listener.}

### 3.2 The Simulated Development

The development of Elspeth in Scottish and Liz in Southern English is shown in Figure 4. Elspeth gradually improves in distinguishing /ɪ/ from /i/. It can be shown that the ratio of the duration reliance and the spectral reliance (in terms of the F1 and duration ranges, respectively) is a good estimate of the slope of the boundary line (cf. §2.1). Ultimately, Elspeth's duration/spectral reliance ratio (the slope of the boundary line in Elspeth's fourth picture) becomes (8.4% · log₂ (500/260)) / (92.8% · log₂ (120/50)) = 0.068 oct/dur. doubling.

The development of Liz in Southern England is very different. Figure 4 shows that her final duration/spectral reliance ratio is 1.04 oct/dur. doubling. The simulated reliance ratios of 0.068 and 1.04 compare well with the optimal ones (§2.2) of 0.075 and 0.98 (the small differences are due to the finite accuracy of the learning process). More details about the evaluation procedure are reported in Escudero (2001).
Fig. 5: Reliance on spectral and duration cues for average real Scottish English listeners (left) and Southern English listeners (right).

If we compare the boundary line of the real Scots (Figure 5) with that of Elspeth (Figure 4), we see that their heights are equal (around 400 Hz) and that their slopes are almost equal (0.050 vs. 0.068 oct/dur.doubling). The real Southerners, by contrast, are quite different from Liz: their category boundary line is much lower (though higher than that of the Scots). This difference could be due to any of the following or more:

(a) In the listening experiment, the spectral cue for the Southerners was enhanced in an unnatural way, i.e. the F1 range in Figure 5 was much larger than their native height contrast. This may have enhanced these listeners' awareness of this cue and thus selectively reduced the duration/spectral reliance ratio for the Southerners only (note that a similar argument is not valid for the Scots, for whom the large F1 range is closer to their native height contrast).

(b) The listening experiment had two properties that may have contributed to lower duration/spectral reliance ratios: (1) the first cue available was spectral, and (2) with isolated vowels, listeners can hardly normalize away the influence of speaking rate on duration, whereas they can partly normalize away the influence of vocal tract size on the basis of the available pitch.

(c) The simulated reliance ratios are sensitive to the standard deviations used for simulating the variation in F1 and duration, but we do not know these values. The simulations are based on the standard deviations estimated in the production experiment (§2.1).

(d) The simulated reliance ratios also depend on the number of tokens in the production experiment. The number of tokens is 100 in Figure 4 and 1 for Figure 5.

(e) The simulated reliance ratios also depend on the number of tokens in the production experiment. The number of tokens is 100 in Figure 4 and 1 for Figure 5.

(f) In general, real listeners have more fine-grained formal models than simulated listeners, and their perception strategies tend to converge with multiple interactions, so their reliance ratios are likely to be more fine-grained and less sensitive to the number of tokens.

(g) The simulated reliance ratios also depend on the number of tokens in the production experiment. The number of tokens is 100 in Figure 4 and 1 for Figure 5.

(h) The simulated reliance ratios also depend on the number of tokens in the production experiment. The number of tokens is 100 in Figure 4 and 1 for Figure 5.

(i) The simulated reliance ratios also depend on the number of tokens in the production experiment. The number of tokens is 100 in Figure 4 and 1 for Figure 5.

(j) The simulated reliance ratios also depend on the number of tokens in the production experiment. The number of tokens is 100 in Figure 4 and 1 for Figure 5.

(k) The simulated reliance ratios also depend on the number of tokens in the production experiment. The number of tokens is 100 in Figure 4 and 1 for Figure 5.

(l) The simulated reliance ratios also depend on the number of tokens in the production experiment. The number of tokens is 100 in Figure 4 and 1 for Figure 5.

(m) The simulated reliance ratios also depend on the number of tokens in the production experiment. The number of tokens is 100 in Figure 4 and 1 for Figure 5.

(n) The simulated reliance ratios also depend on the number of tokens in the production experiment. The number of tokens is 100 in Figure 4 and 1 for Figure 5.

(o) The simulated reliance ratios also depend on the number of tokens in the production experiment. The number of tokens is 100 in Figure 4 and 1 for Figure 5.

(p) The simulated reliance ratios also depend on the number of tokens in the production experiment. The number of tokens is 100 in Figure 4 and 1 for Figure 5.

(q) The simulated reliance ratios also depend on the number of tokens in the production experiment. The number of tokens is 100 in Figure 4 and 1 for Figure 5.

(r) The simulated reliance ratios also depend on the number of tokens in the production experiment. The number of tokens is 100 in Figure 4 and 1 for Figure 5.

(s) The simulated reliance ratios also depend on the number of tokens in the production experiment. The number of tokens is 100 in Figure 4 and 1 for Figure 5.

(t) The simulated reliance ratios also depend on the number of tokens in the production experiment. The number of tokens is 100 in Figure 4 and 1 for Figure 5.

(u) The simulated reliance ratios also depend on the number of tokens in the production experiment. The number of tokens is 100 in Figure 4 and 1 for Figure 5.

(v) The simulated reliance ratios also depend on the number of tokens in the production experiment. The number of tokens is 100 in Figure 4 and 1 for Figure 5.

(w) The simulated reliance ratios also depend on the number of tokens in the production experiment. The number of tokens is 100 in Figure 4 and 1 for Figure 5.

(x) The simulated reliance ratios also depend on the number of tokens in the production experiment. The number of tokens is 100 in Figure 4 and 1 for Figure 5.

(y) The simulated reliance ratios also depend on the number of tokens in the production experiment. The number of tokens is 100 in Figure 4 and 1 for Figure 5.

(z) The simulated reliance ratios also depend on the number of tokens in the production experiment. The number of tokens is 100 in Figure 4 and 1 for Figure 5.
Discussion

We have hypothesized that adult listeners have a perception tuned accurately to their production environment, and we have proposed a model for the knowledge behind this near-optimal perception and its development. Our simulations show that our model indeed implements a near-optimal integration of two acoustic cues (i.e., cue reliance depends on cue reliability) and handles its development successfully. The model is built on Optimality Theory rather than other possible frameworks in order that our model becomes part of phonological theory.

Future research will have to model category split and/or merger and the influences of consonant voicing, the number of syllables, stress, speaking rate, inter-speaker variation, and dialect interactions. Future work involves second-language perception as well as longitudinal studies.

References


