This paper is to appear, with different page numbers, in

Proceedings of the 25th Penn Linguistics Colloquium
(*Penn Working Papers in Linguistics*)

Its size had to be limited to 14 pages.

April 26, 2001

Modelling the Perceptual Development of Phonological Contrasts with Optimality Theory and the Gradual Learning Algorithm

Paola Escudero and Paul Boersma

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 acoustically detailed input to discrete phonological categories. As an example, this paper considers the development of the perception of the English vowels /ɪ/ 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, /ɪ/ and /i/ differ much more in F1 and much less in duration than for Southern English speakers.

In this paper, we hypothesize that humans have an optimal perception strategy that minimizes the probability of confusion and that there is a knowledge that underlies the implementation of this strategy. We model the knowledge behind speech perception as an Optimality-Theoretic perception grammar, and we model the acquisition of this knowledge with the Gradual Learning Algorithm. Using an environment based on real production data, we simulate the development of a Scottish and a Southern English listener, and show that the Scot comes to rely almost exclusively on height (F1) when distinguishing /ɪ/ and /i/, whereas the Southerner comes to rely on both height and duration, and so the model indeed implements an optimal strategy for acoustic cue integration. Perception experiments show that real Scots and real Southerners also use this optimal strategy in their own environments.

We find, therefore, that perceptual strategies depend on the production environment, and that we can successfully model this dependency within the framework of stochastic Optimality Theory, thus bringing speech-processing systems within the reach of formal phonological theory.

1 Production of /ɪ/ and /i/ in Scottish and Southern English

Our explanation and modelling of the accurate perception of acoustic detail (§2, §3) requires that we accurately measure how the two vowels /ɪ/ and /i/ are realized in Scotland and in the South of England.

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 obtain some realistic variation with respect to 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 sentences, in order to see if he understood the task. The speaker was then asked to handle five decks of 50 cards each, and after a break he was asked to handle the remaining five decks. The experimenter was sitting beside the speaker with a copy of the word list, on which she or he could mark 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 and in Figure 1. We use geometric averaging for F1 as well as for duration, because both dimensions have only positive values (so that effect sizes and standard deviations tend to be constant along a logarithmic scale). The standard deviations 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 category is a minor factor in determining the duration value (which depends much more on the number of syllables and on the voicing of the following consonant), whereas it is the primary factor for the Southern English speaker.

	μ	σ		μ	σ		μ	σ		μ	σ
ship	90.7	0.183	*lid*	134.0	0.182	*Sni-*	55.5	0.151	*fil-*	76.8	0.096
	480	0.038		480	0.051	*cker*	489	0.098	*ling*	492	0.054
sheep	92.0	0.143	*lead*	162.2	0.184	*snea-*	56.2	0.194	*fee-*	93.1	0.095
	327	0.067		324	0.064	*ker*	378	0.059	*ling*	346	0.034

Table 1: Geometric averages (μ) of duration (top) and F1 (bottom), expressed in ms and Hz, and their standard deviations (σ), expressed in duration doublings and octaves, for the Scottish English speaker.

	μ	σ		μ	σ		μ	σ		μ	σ
ship	55.7	0.176	*lid*	75.0	0.128	*Sni-*	48.0	0.155	*fil-*	63.2	0.168
	331	0.057		359	0.086	*cker*	287	0.101	*ling*	379	0.069
sheep	103.1	0.125	*lead*	120.3	0.111	*snea-*	91.4	0.101	*fee-*	105.4	0.159
	287	0.085		290	0.098	*ker*	278	0.095	*ling*	313	0.086

Table 2: The same for the Southern English speaker.

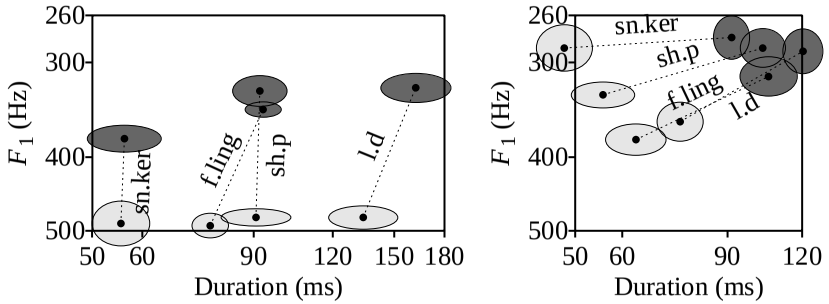

Fig. 1: Scottish (left) and Southern (right) production of /ɪ/ (light) and /i/ (dark). The axes are logarithmic. The ellipses show the standard deviations.

1.3 Relative Cue Use

Our modelling of the perception of the /ɪ/-/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.

Scottish		μ	σ	Southern		μ	σ
/ɪ/	dur.	84.8 ms	0.485	/ɪ/	dur.	59.7 ms	0.284
	F1	485 Hz	0.066		F1	337 Hz	0.170
/i/	dur.	94.0 ms	0.565	/i/	dur.	104.6 ms	0.188
	F1	343 Hz	0.105		F1	292 Hz	0.110

Table 3: Duration and F1 for /ɪ/ and /i/ for the Scottish and Southern speaker, averaged across the four contexts, and the total standard deviations.

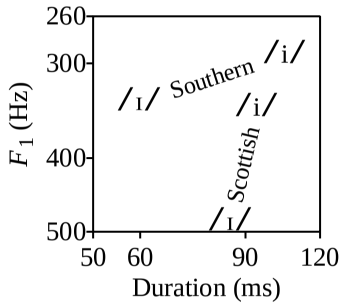

Fig. 2: Duration and F1 for /ɪ/ and /i/ for the Scottish and Southern English speaker, averaged across the four contexts.

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 consonantal context (i.e. whether or not the following consonant is voiced) and 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 /ɪ/ 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** oct/dur.doubling. This number is equal to the slope of the imaginary line that connects the Scottish /ɪ/ 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** oct/dur.doubling. Apparently, the Scot prefers the F1 dimension (or disfavours the duration dimension) 13 times more than the Southerner does.

2 Modelling the Perception Process and its Acquisition

In general, the perception process maps multiple acoustic cues to multiple phonological contrasts simultaneously (e.g. vowel duration plays a role in the perception of the vowel contrast as well as in the perception of the voicing of the following consonant). This paper will restrict itself to the integration of *two* acoustic cues into *one* phonological contrast.

This section presents our model of perceptual development, illustrated by the behaviour of virtual Elspeth and virtual Liz, who grow up in virtual Scottish and Southern English environments, respectively. We will show how the perceptual strategy implemented by the model depends on the reliability of the two cues in the virtual production environments. In §3, we will verify the predictions of this model in a computer simulation and show that the predictions are borne out by the behaviour of real listeners.

2.1 The Virtual Production Environment

We assume that the vowels that Elspeth and Liz hear are drawn from Gaussian distributions that are centred about the mean F1 and duration values for the Scottish and Southern English speakers (Table 3), so that the relative cue use in Elspeth's and Liz' production environments is equal to that of the real speakers in §1.3.

For our Gaussian production distributions, we choose fixed standard deviations of $\sigma_{F1} = 0.20$ octaves and $\sigma_{dur} = 0.40$ duration doublings for both vowels and both dialects. These values are different from those in Table 3, for the following reasons. The standard deviations in Table 3 include the variation that no listener can normalize for, i.e. the random variation between tokens of the same utterance, as well as the variation due to the consonant environment and the number of syllables, which listeners can partially normalize away (this would lead to lower σ than those in Table 3). However, variations due to speaking rate, stress, and vocal tract size were not included in our production experiment and will have unknown but positive effects on the variation in the listener's input data (this would lead to higher σ values). In the production experiment (Tables 1, 2, and 3), we saw that the standard deviations for duration tended to be higher than those for F1, so we use the somewhat arbitrary values of 0.20 and 0.40. These values are large enough to ensure that a wide range of duration-F1 pairs will occur in Elspeth's and Liz' environments. Unfortunately, the results of our simulations will be very sensitive to the exact values of these standard deviations.

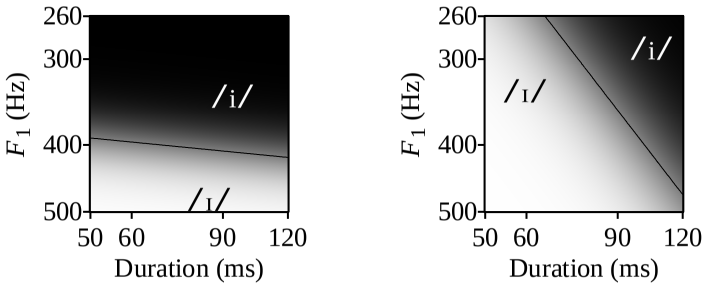

Fig. 3: The production environments for virtual Scottish Elspeth (left) and Southern English Liz (right).

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 /ɪ/-/i/ contrast. The cues can be more or less reliable according to how much information they give to the perception process, e.g. the reliability of the F1 cue for Scottish depends on how little the F1 values for /ɪ/ overlap with those for /i/. The numeric measure expresses the cue ranges (§1.3) in terms of standard deviations. For Scottish, the F1 range of 0.500 octaves is equivalent to 0.500/0.20 = 2.5 σ_{F1} (very reliable), the duration range of 0.149 doublings amounts to 0.37 σ_{dur} (unreliable). For Southern English, the F1 range is 1.04 σ_{F1} (not very reliable), the duration range 2.01 σ_{dur} (quite reliable). From these values, we can predict that an ideal Elspeth, who will rely mainly on reliable cues, will rely almost exclusively on F1 and hardly on duration, whereas Liz will rely on duration primarily, on F1 secondarily.

Figure 3 shows how likely it is that any duration-F1 pair was intended as /i/ (black) or as /ɪ/ (white) in the two dialects, assuming that the two vowels occur equally often in the environment. The black line connects the duration-F1 pairs that are equally likely to be /i/ or /ɪ/. It can be shown that the slope of this equal-likelihood line is given by σ_{F1}/σ_{dur} times the ratio of the reliabilities. For the virtual Scottish environment, the equal-likelihood slope is (0.20/0.40) · (0.37/2.5) = **0.075** oct/dur.doubling, for the Southerner it is **0.98** oct/dur.doubling, which is (again) 13 times greater than the Scottish.

2.2 The Optimal Perception

We hypothesize that listeners minimize the probability of miscomprehension 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.

Suppose, for instance, that both Scottish Elspeth and Southern English Liz are confronted with the same acoustic event (duration-F1 pair), for example [349 Hz, 74 ms]. Figure 3 shows that if the two listeners exhibit optimal perception (i.e. if they manifest maximum-likelihood behaviour), then Elspeth will perceive this acoustic event as /i/, and Liz will perceive the same event as /ɪ/. More generally, they will perceive everything above their own equal-likelihood line as /i/, everything below as /ɪ/.

The optimal perceiver will therefore have a decision boundary in perception that coincides exactly with the equal-likelihood line in her production environment. The slope of this category boundary is a measure of the ratio of the listener's reliance on duration and her reliance on F1. The optimal duration/spectral reliance ratio for the Scottish listener, therefore, is **0.075** oct/dur.doubling, for the Southern listener it is **0.98** oct/dur.doubling, again 13 times as high as that of the Scot. Taking into account the slope formula at the end of §2.1, we see that such duration/spectral reliance ratios depend directly on the reliability of the cues in the production environment.

We could now test our hypothesis against real listening experiments. However, we believe that there is a need to explain in detail the knowledge that underlies overt perceptual behaviour. Therefore, the next two sections will present a model that answers the questions: how do listeners implement an optimal perception strategy, and how do they *learn* to do it? We will later (§3) verify the validity of our model and test whether the model is realistic.

2.3 Modelling the Perception Process

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 with rankings that choose an optimal output (here: phonological category) on the basis of an input (here: acoustic event). The decision scheme works according to the framework of Optimality Theory (OT; Prince & Smolensky 1993), or more specifically its probabilistic version (stochastic OT; Boersma 1998).

Boersma (1998:164) proposed constraints for mapping acoustic cues to phonological categories. In the case at hand, we label the categories arbitrarily as /ɪ/ and /i/. We divide the F1 continuum arbitrarily into 21 logarithmically equal steps, giving constraints from "260 Hz should not be perceived as /ɪ/" to "500 Hz should not be perceived as /ɪ/", and analogously for /i/. We also divide the duration continuum into 21 steps,

giving constraints from “50 ms should not be perceived as /ɪ/” to “120 ms should not be perceived as /ɪ/” (and the same for /i/). So we use 84 negatively worded constraints for modelling the perception of the two vowels (using positively worded constraints such as “260 Hz should be perceived as /i/” would not work if we had more than two categories, e.g. if we also wanted to take into account the vowels /e/ and /ɛ/).

The underlying knowledge of Elspeth’s perception of the acoustic event [349 Hz, 74 ms] can now be represented as the constraint ranking in Tableau 1. Only four of our 84 constraints are relevant here. The highest ranked of these must be “349 Hz is not /ɪ/”, because of the large distance (in terms of standard deviations) between 349 Hz and the mean F1 for /ɪ/ (§2.2). Only the two relevant vowel categories are shown as candidates in Tableau 1. When the acoustic event [349 Hz, 74 ms] arrives, the tableau will select the candidate /i/ as the winner (i.e. as the actually perceived category) because this candidate violates the least high-ranked constraints.

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

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

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

Tableau 2: The perception of the acoustic event [349 Hz, 74 ms] for Liz, who lives in a Southern English production environment.

The knowledge underlying the perception of the same acoustic event for Liz is shown in Tableau 2. Her two F1 constraints are ranked in the reverse order from Elspeth’s, and she will choose to perceive /ɪ/.

We should note that in stochastic OT, the listener has no direct knowledge of probabilities. Her only knowledge resides in the rankings of the constraints, and any apparent optimal behaviour is derived from that.

2.4 Modelling the Acquisition of Perception

It’s fine to have those rankings, but how did they come about? Are Elspeth and Liz able to learn this optimal strategy at any point during their lives? Our answer is that they succeed by applying the Gradual Learning Algorithm (Boersma and Hayes 2001) to the perception grammar.

For example, Elspeth may entertain at a certain point during her perceptual development a grammar that would be appropriate for Liz. As a consequence, Elspeth perceives [ʃɐ̯p], with the vowel cues [349 Hz, 74 ms], as /ɪ/, as shown in Tableau 3. However, her environment is Scottish, so this acoustic event is much more likely to have been related to the fluffy animal (underlyingly |ʃip|) than to the floating means of transportation (|ʃɪp|).

[349 Hz, 74 ms]	349 Hz not /i/	74 ms not /i/	74 ms not /ɪ/	349 Hz not /ɪ/
☞ /ʃɪp/			←*	←*
√ /ʃip/	*!→	*→		

Tableau 3: Error-driven learning by the Gradual Learning Algorithm in an Optimality-Theoretic perception grammar.

If we assume that Elspeth detects her perception error (because the semantic context tells her that |ʃip| ‘sheep’ is the correct recognition of this particular /ʃɪp/ perception), she will change her perception grammar by raising the rankings of all the constraints violated in her incorrect winner and by lowering the rankings of all the constraints violated in the form that she considers correct, thus increasing the probability that she will perceive [349 Hz, 74 ms] as /ʃip/ on the next occasion. The rankings are changed by only a small step (called the *plasticity*) along the continuous ranking scale of stochastic OT, but after a large number of perception errors the rankings of the constraints will have become that of an adult Elspeth, as in Tableau 1.

We should note that the Gradual Learning Algorithm has no knowledge of any optimal perception strategy. Boersma (1998:338) nevertheless showed that in the case of single-cue categorization, this algorithm leads to a *probability-matching* perceiver, i.e. one whose category boundaries coincide with the equal-likelihood boundaries of production, but whose boundary slopes are smooth (as in real listeners) rather than sudden (as for an ‘optimal’ maximum-likelihood perceiver). The next section will tell us whether this desirable near-optimal property of the algorithm extends to the two-cue case.

3 The Simulations

We will simulate here the perceptual development of our virtual listeners Elspeth and Liz, who we introduced in §2, from infancy through adulthood. We will test whether they acquire an optimal perception (i.e. whether our model indeed implements a maximum-likelihood type of behaviour) and we will compare their final stages with those of real adult Scottish and Southern English listeners.

In Elspeth's and Liz' initial state, all 84 constraints are ranked at the same height, so that the baby is equally likely to perceive any acoustic event as /ɪ/ or as /i/. We understand that this is a rather artificial initial state. It assumes that the baby has different lexical representations for /ɪ/ and /i/ without being able to distinguish them perceptually yet. In reality, the emergence of lexical categories must be based on a perceptual distinction. However, this paper will not pursue a discussion of category emergence.

3.1 The Simulated Development

We simulated the development of a Scottish and a Southern English listener by feeding them with input-output pairs drawn randomly from the Gaussian distributions (§2.1) that represent the probability of occurrence of each of the 441 F1-duration values for /i/ and /ɪ/ in the learner's environment. Throughout her life, each listener receives 1000 data per month, and changes some constraint rankings every time she notices a mismatch between her perceived category and the correct lexical category. During the first 10 virtual months, the plasticity is 1.0, which means that the rankings are changed by an amount of 1.0 along the ranking scale (this amount is one half of the evaluation noise of stochastic OT, which we keep constant at 2.0 throughout our simulations). Between 10 and 100 virtual months of age, her plasticity is only 0.1, which means that she learns more slowly, but also more accurately because the evaluation noise is still 2.0. Between 100 and 1000 months, her plasticity is only 0.01.

Figure 4 shows the perceptual performance of Elspeth in various stages. For each picture, we measured Elspeth's output distribution by confronting her with 1000 instances of each of the 441 F1-duration pairs, and counting the number of /ɪ/ and /i/ responses for each of these 441 possible acoustic events. Black areas stand for /i/ perceptions, white areas for /ɪ/ perceptions, and the black curve is the 50% contour, i.e. the category boundary 'line'. The spectral reliance ("spec.rel.") is computed as the average fraction of /i/

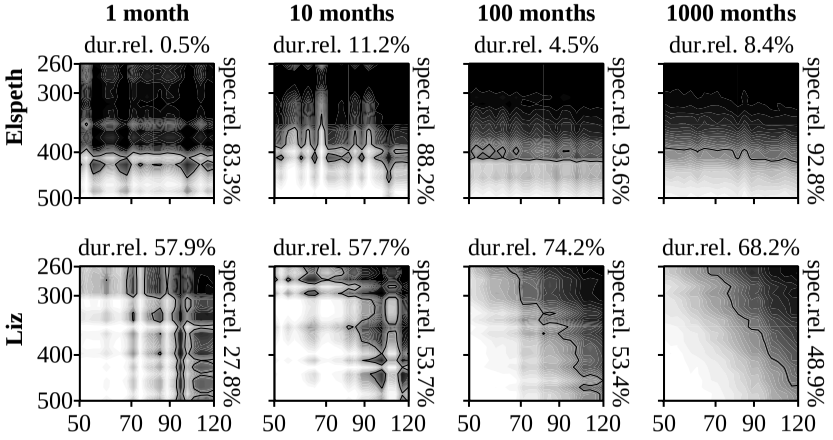

Fig. 4: The perceptual development of Scottish Elspeth and Southern Liz.

responses along the top edge minus the average fraction of /i/ responses along the bottom edge of the picture. The duration reliance ("dur.rel.") is likewise computed from the fractions of /i/ responses along the right and left edges (Escudero 2001).

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 $(8.4\% \cdot \log_2 (500/260)) / (92.8\% \cdot \log_2 (120/50)) =$ **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 generally, the final stages in Figure 4 are very similar to Figure 3. We conclude that our model indeed implements a maximum-likelihood-like (probability matching) listener, even when confronted with multiple cues.

A dimensionless *language-specific* reliance ratio can be computed by normalizing the duration/spectral reliance ratios for the cue ranges in production (§1.3). For Elspeth, this gives a language-specific reliance ratio of

0.068·(0.149/0.500) = **0.020**, i.e. she relies on the spectral cue 50 times more than on the duration cue when listening to the contrast between Scottish /ɪ/ and Scottish /i/. Liz has a language-specific reliance ratio of 1.04·(0.809/0.207) = **4.1**, i.e. she relies 4.1 times more on duration than on F1 for distinguishing Southern /ɪ/ and /i/.

3.2 Comparison with Real Listeners

We can now test the optimal-perception hypothesis by comparing the results of the simulations with those of an older experiment with real listeners, reported in Escudero (2001). Figure 5 shows the average cue reliance of 20 Scottish English listeners, and that of 21 Southern English listeners, all of whom were tested with the same large duration-F1 stimulus continuum of synthetic vowels (F2 was also varied). The duration/spectral reliance ratio (i.e. an estimate of the boundary slope in Figure 5) for the Scots is (10.6%·log(480/344)) / (93.4%·log(177/83)) = **0.050** oct/dur. doubling, and for the Southerners it is **0.233** oct/dur.doubling.

If the average cue values in the listeners' language environments are equal to those that we measured in our production experiment, the language-specific duration/spectral reliance ratios (§3.1) can be computed as **0.015** for the Scots (i.e. they rely on F1 70 times as much as on duration) and as **0.93** for the Southerners (i.e. they rely equally on F1 and duration).

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 for the Scots) and the slope is much smaller (0.233 vs. 1.04; §3.1), though much greater than that of the Scots. This difference could be due to any or all 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 duration range in Figure 5 corresponds to their own natural, though allophonic, variation in duration). This could be solved by testing listeners with stimulus sets that do not extend beyond the duration and F1 ranges that are appropriate for their dialect;

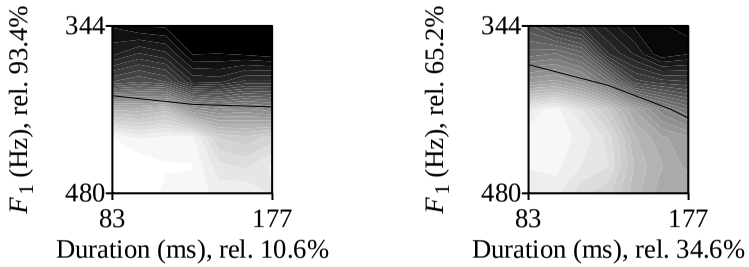

Fig. 5: Reliance on spectral and duration cues for average real Scottish English listeners (left) and Southern English listeners (right).

(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 (§2.1) used for simulating the variation in F1 and duration, but we do not know what these are, since we do not know to what extent the listener compensates for consonant environment, number of syllables, stress, or speaking rate. If we double the Southern σ_{dur} to 0.80 doublings (or halve the Southern σ_{F1} to 0.10 octaves), Liz will acquire a duration/spectral reliance ratio of about 0.25 oct/dur.doubling, i.e. equal to that of the real listeners;

(d) In general, real listeners have contact with multiple dialects, so their perception strategies tend to converge, whereas the simulated listeners were raised in completely isolated environments;

(e) The Southern English speaker destressed the target words, as expected in the environment "THIS is a __ as well", but the Scottish speaker gave the target word a secondary stress. The effect of this remains unknown to us.

(f) The Southern speaker may not have been representative of the environment of the Southern listeners.

Most of these facts seem to support our view of the optimal perceiver, whose more fine-grained formal modelling, however, has to await future research.

4 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 for its acquisition. We model the perception process with an Optimality-Theoretic constraint grammar that maps raw acoustic input to discrete phonological categories, and we model the acquisition of this process with the Gradual Learning Algorithm, which reranks the constraints in case of misclassification.

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. In real listeners, differences in the production environment turn out to lead to similar differences in perception. So we can conclude that these listeners have a grammar similar to the one proposed in our model. We use 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

Boersma, Paul. 1998. *Functional phonology*. PhD dissertation, University of Amsterdam. The Hague: Holland Academic Graphics.

Boersma, Paul, and Bruce Hayes. 2001. Empirical tests of the Gradual Learning Algorithm. *Linguistic Inquiry* 21, 45-86.

Escudero, Paola. 2001. The role of the input in the development of L1 and L2 sound contrasts: Language-specific cue weighting for vowels. In *Proceedings of the 25th Boston University Conference on Language Development, 250–261*. Somerville, Mass.: Cascadilla Press.

Prince, Alan, and Paul Smolensky. 1993. *Optimality Theory: Constraint interaction in generative grammar*. Technical Report TR-**2**, Rutgers University Center for Cognitive Science.

Escudero:
School of Linguistics & Applied Language Studies
The University of Reading
Whiteknights, PO Box 218
Reading, RG6 6AA, England
p.r.escudero@reading.ac.uk

Boersma:
Institute of Phonetic Sciences
University of Amsterdam
Herengracht 338
1016 CG Amsterdam, The Netherlands
paul.boersma@hum.uva.nl