Learning Abstract Phonological from Auditory Phonetic Categories: An Integrated Model for the Acquisition of Language-Specific Sound Categories

Paul Boersma*, Paola Escudero† and Rachel Hayes‡
*University of Amsterdam, Amsterdam, The Netherlands
†Utrecht University, Utrecht, The Netherlands
‡University of Arizona, Tucson, USA
E-mail: paul.boersma@hum.uva.nl, paola.escudero@let.uu.nl, rhayes@u.arizona.edu

ABSTRACT

We introduce a two-stage model for the perceptual acquisition of speech sound categories within the framework of Stochastic Optimality Theory and the Gradual Learning Algorithm [1]. During the first stage, learning of language-specific sound categories by infants is driven by distributional evidence in the linguistic input. This *auditory-driven learning* leads to a warping of the baby's perceptual space, to discrimination curves, to the perceptual magnet effect, and ultimately to the creation of phonetic categories. In the transition to the second stage, these phonetic categories turn into simple abstract phonological categories. During the second stage, when the lexicon is in place, *lexically-driven learning* will develop more abstract representations and optimize multi-dimensional perception. The results of our simulations compare well with findings from the infant literature and from cross-language studies.

1. INTRODUCTION

Infants have a remarkable capacity to calculate the statistical distributions of auditory phonetic information in their linguistic input. It has been argued that their knowledge of these distributions ultimately leads to the creation of phonetic categories at 6-8 months of age [2]. This *auditory-driven learning* has been modelled in the domains of cognitive science and psychology with neural networks whose outcomes automatically reflect the statistical distributions of the language input [3]. Once the lexicon is in place, however, more abstract levels of representation come into being: features combine into segments (as witnessed by the development of the weighting of auditory cues [4]), and allophones combine into phonemes [5]. For cue weighting, this *lexicon-driven learning* has been modelled in the domain of linguistics by Stochastic Optimality Theory and the Gradual Learning Algorithm [6]. The present work proposes an underlying mechanism common to both kinds of learning, and explicitly models the transition between the two. The model employs a gradual perceptual learning device that is fed by two types of evidence: (i) acoustic events in the linguistic input, which give birth to 'phonetic' categories; and (ii) lexical representations, which lead to the development of 'phonological' categories.

2. LEARNING PHONETIC CATEGORIES

2.1 Speech perception in Optimality Theory

We model early speech perception learning according to the proposal of Functional Phonology [7], where three families of competing constraints determine the mapping from auditory inputs to phonetic categories. A family of PERCEIVE constraints militates against not perceiving auditory inputs at all. Thus, PERCEIVE (F1: [700 Hz]) requires the listener to treat an auditory input with an F1 value of 700 Hz as a member of *some* category. The particular category that it is assigned to is determined by two types of constraints: *CATEG(ORIZE) and *WARP. The *CATEG family punishes perceptual categories with particular acoustic values, e.g., *CATEG (F1: /700 Hz/) militates against perceiving an incoming F1 into the 'category' /700 Hz/. The *WARP family requires every acoustic input to be perceived as a member of the *most similar* available category. Thus, *WARP (F1: 40 Hz) says that an acoustic input with an F1 of 680 Hz should not be perceived as any F1 'category' that is 40 Hz off (or more), i.e. as /640 Hz/ or /720 Hz/ or anything even farther away.

2.2 The initial state

We propose that in the initial state of the infant, all *CATEG constraints are ranked high, and all PERCEIVE constraints are ranked low. This means that it is worse for the child to perceive an incoming F1 as something than to not perceive the incoming F1 at all:

[340 Hz]	*CATEG (/320/)	*CATEG (/340/)	PERCEIVE ([340])	*WARP (20)
/320 Hz/	*!			*
/340 Hz/		*!		
☞ /–/			*	

This tableau (which for reasons of space contains only a very small subset of all the constraints and candidates) shows that an incoming [340 Hz] will be perceived as the null candidate /–/, which violates the constraint PERCEIVE (F1: [340 Hz]), because the competing candidates /320 Hz/ and /340 Hz/ violate higher-ranked constraints. *WARP (F1: 20 Hz) is violated if [340 Hz] is perceived as /320 Hz/; its ranking does not contribute to determining the winner here.

2.3 The learning mechanism

The learner will not be satisfied with always perceiving the null category. We propose that when she hears an F1 of [340 Hz], her blind innate distributional learning device will tell her that she should have perceived this as the 'identical' value /340 Hz/. This value should therefore be included in the tableau's top left cell, which contains all the information given to the learner. The child will now consider one of her candidates 'correct', as shown with a check mark in the next tableau:

[340 Hz] /340 Hz/	*CATEG (/320/)	*CATEG (/340/)	PERCEIVE ([340])	*WARP (20)
/320 Hz/	*!			*
√ /340 Hz/		*!→		
☞ /–/			←*	

Now that the learner knows that she has made an error, she can *learn*: she will lower the ranking of the constraints violated in the form that she considers correct (this is shown by the rightward arrow in the tableau) and raise the ranking of the constraints violated in her own winning form (the leftward arrow). This procedure is the Gradual Learning Algorithm [1]. In Stochastic Optimality Theory, constraints are ranked along a continuous scale, and the algorithm typically reranks them in small steps, achieving accuracy and robustness. After many incoming [340 Hz] values, *CATEG (F1: /340 Hz/) will ultimately fall below PERCEIVE (F1: [340 Hz]), and the infant will perceive this input 'correctly':

[340 Hz] /340 Hz/	*CATEG (/320/)	PERCEIVE ([340])	*CATEG (/340/)	*WARP (20)
/320 Hz/	*!			*
√ ☞ /340 Hz/			*	
/–/		*!		

Learning will now stop, since the learner now considers the output of her grammar correct...

But the child does not only hear [340 Hz] values. Suppose that she will also hear some F1 values of [320 Hz], but less often than [340 Hz]. The constraints *CATEG (F1: /320 Hz/) and PERCEIVE (F1: [320 Hz]) will move, but less so than the constraints for 340 Hz. A possible ranking is shown in the following tableau:

[320 Hz] /320/	*WRP 60	PERC [340]	PERC [320]	*CAT /320/	*CAT /340/	*WRP 20
√ /320/				*!→		
☞ /340/					←*	←*
/–/			*!			

We see how the auditory input [320 Hz] is perceived into the 'category' /340 Hz/. This happens because the *WARP constraint against perceiving an input into a category that is off by 20 Hz is ranked very low (20 Hz is below the just noticeable difference for formants [8]).

But incoming F1 values that are more distant from 340 Hz will not be perceived as /340 Hz/. See next tableau.

[280 Hz] /280/	*WRP 60	PERC [280]	*CAT /280/	*CAT /320/	*CAT /340/	*WRP 40
√ /280/			*!→			
☞ /320/				←*		←*
/340/	*!				*	*
/–/		*!				

Thus, if [280 Hz] is even less common in the input than [320 Hz], an incoming [280] will be perceived as /320/. The listener has established a compromise: the high ranking of *WARP (60) tells her that /340 Hz/ is too far off, and the relatively high ranking of *CATEG (/280 Hz/) tells her that /280 Hz/ is a too uncommon 'category'. We observe that as a result of distributional skewings in her language environment, the infant will warp her perceptual space in favour of the commonest F1 values. A situation in which some F1 values are more common than others is likely to occur in practice, namely as the result of a finite number of vowel height categories in the speakers of the ambient language. Suppose a language has vowels with average produced heights of 340 and 480 Hz. If we shelve the problems of between-speaker variation, the environment will have an F1 distribution with peaks around [340] and [480] Hz. The model just described predicts that the infant will learn to map incoming F1 values in the following way:

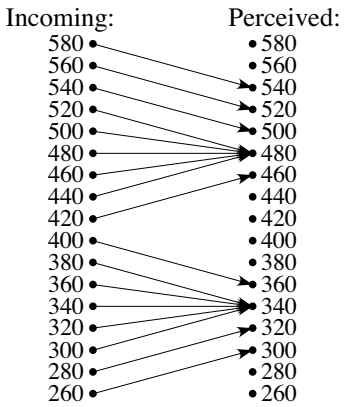

Figure 1: Warping of the perceptual space.

2.4 Simulation of distributional learning

The previous two tableaux show that in the later stages a learning step no longer simply involves one *CATEG lowering and one PERCEIVE raising, but involves an intricate movement of *CATEG constraints. To check if Figure 1 actually results, we ran a computer simulation (in Praat) on a 20-Hz discretization of the F1 continuum, giving 122 constraints. Their initial rankings were:

*WARP (F1: 800 Hz): ranked at a height of 800
*WARP (F1: 780 Hz): ranked at a height of 780
...
*WARP (F1: 60 Hz): ranked at a height of 60
*CATEG (F1: /200 Hz/), *CATEG (/220 Hz/), ..., *CATEG (/1000 Hz/): all ranked at a height of 0
PERCEIVE (F1: [200 Hz]), PERCEIVE ([220 Hz]), ..., PERCEIVE (F1: [1000 Hz]): all ranked at -1000
*WARP (F1: 40 Hz): ranked at -10^9
*WARP (F1: 20 Hz): ranked at -10^9

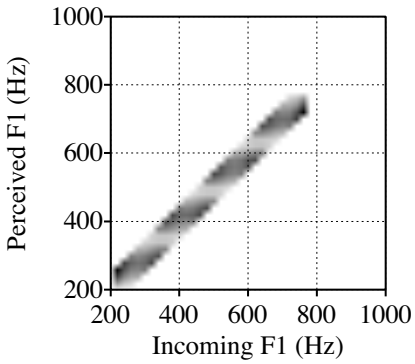

Figure 2: Noisy warping of the perceptual space.

We fed the learner with 400,000 F1 values drawn from an environment with four vowels with average F1 values of 280, 420, 560, and 700 Hz and standard deviations of 30 Hz. The *plasticity* (the size of the step by which rankings can change on each input) was taken to drop gradually from 1.0 to 0.001, and the *evaluation noise* (power of the noise added to the ranking of each constraint at evaluation time) was taken constant at 2.0. After learning finished, we ran 1000 tokens of each of the 41 possible F1 values through the resulting grammar. Figure 2 shows us how the listener perceived each F1 value. We see that as a result of the noisy evaluation, every incoming F1 value can be perceived as several different F1 values. Nevertheless, there is a clear warping of the perceptual space: perceived values cluster around 280, 420, 560, and 700 Hz. Inputs higher than 800 Hz are still perceived as /–/ because of their rarity.

2.5 The discrimination task

The warping of the auditory F1 space leads to a change in perceptual distances. Around the distributional peak of 420 Hz, for instance, distances shrink by one third, since [390] is on average perceived as /400/, and [450] as /440/. Around the distributional valley of 490 Hz, the reverse happens: [460] is perceived on average as /445/, [520] as /535/. Figure 3 shows the perceived distance for an acoustic difference of 60 Hz centred around every acoustic F1 value. Thus, the perceptual space is warped in such a way that differences near ambient category centres are less well perceived than differences near ambient category boundaries: we observe discrimination effects **without the infant having any discrete categories yet**. This *perceptual magnet* effect occurs in real infants [9] and has been modelled with simulated neural networks [10]. We have been able to model the same effect within a linguistic framework.

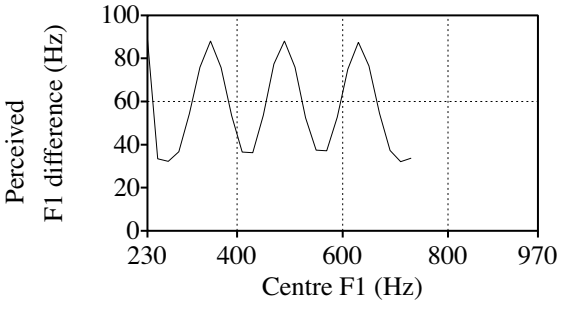

Figure 3: Perceived distances in the warped perception of F1, for an acoustic distance of 60 Hz.

2.6 Conversion to discrete categories

The output of the 'grammar' is commensurate to its input, i.e., the input and output are expressed in the same units, namely Hertz. This special situation allows us to feed the output back to the input, giving a self-enhancing circuit that is capable of warping the frequencies quite far away from their original values. In the example of Figure 1, an input of 580 Hz maps to 540 Hz in the first cycle; in the second cycle, this 540 Hz maps to 500 Hz, which maps to 480 Hz in the third. After this, it will not change any further (480 Hz maps to 480 Hz). Thus, all inputs between 410 and 580 Hz will ultimately map to 480 Hz, and all inputs between 260 and 410 Hz will map to 340 Hz. The number of possible outputs has now become finite, and we can call the values of /340 Hz/ and /480 Hz/ *discrete phonetic categories*.

3. LEARNING PHONOLOGICAL CATEGORIES

Now that discrete categories exist, the child can give them arbitrary labels, severing the connection to the actual continuous F1 values. To stay in line with traditional phonological terminology, we label /340 Hz/ as /high/ and /480 Hz/ as /mid/.

3.1 Lexically-driven optimization of perception

Once categories are discrete labels (feature values), the child can store lexical entries economically as structures consisting only of these labels and their discrete temporal and hierarchical relations. The *WARP constraints have to be translated to accommodate the new categories: if the category centres for /high/ and /mid/ are 340 and 480 Hz, *WARP (F1: 80 Hz) will be split into "[260 Hz] is not /high/", "[420 Hz] is not /high/", "[400 Hz] is not /mid/", and "[560 Hz] is not /mid/", all initially ranked equally high. The universally lower-ranked *WARP (F1: 60 Hz) splits into four lower-ranked constraints, among which "[420 Hz] is not /mid/". These initial rankings cause the listener to have a reasonably good initial categorization performance:

[420 Hz]	PERC [420 Hz]	[420 Hz] is not /high/	[420 Hz] is not /mid/
/high/		*!	
☞ /mid/			*

But the lexicon can now act as a supervisor for achieving a more accurate perception. If it tells the listener that she should have perceived this particular token as /high/ rather than as /mid/, perhaps because the semantic context forces a recognition of *sheep* rather than *ship*, the listener will take appropriate action by making sure that she will be more likely to perceive the next [420 Hz] as /high/ (in the tableau, the lexical recognition is part of the input, i.e. the facts known to the child, and is therefore written between pipes in the top left cell):

[420 Hz] | high |	PERC [420 Hz]	[420 Hz] is not /high/	[420 Hz] is not /mid/
√ /high/		*!→	
☞ /mid/			←*

In the case of noisy evaluation the learner will ultimately become a *probability-matching listener*, i.e., her probability of perceiving [420 Hz] as /high/ or /mid/ will mimick the distribution of underlyingly /high/ and /mid/ realized as [420 Hz] in her environment [7].

3.2 High-level perceptual integration

There will be more discretized continua than just F1. For instance, acoustic vowel duration may have been divided into categories arbitrarily labelled /short/ and /long/, with constraints such as "[91 ms] is not /short/". Around 9 months of age, infants start to integrate multiple categories into higher-level abstractions: 9 month old but not 6 month old infants use both sequential and rhythmic information to recognize two-syllable words in a larger speech stream [11], and multi-dimensional categorization occurs in the development of visual categories by infants from 9 months on as well [12]. Thus, in some varieties of English, the vowel of the lexical entry *sheep* will be stored with the feature values /high, long/, the vowel of *ship* with /mid, short/, and the child will learn to use both F1 and duration in perceiving this /i/–/ɪ/ contrast. This perception can be modelled with initially high ranked constraints against feature co-occurrence, i.e. */high, long/, */high, short/, */mid, long/, and */mid, short/. The tableau at the bottom of this page, which was the end result of our computer simulation of learning with plausibly distributed *sheep–ship* tokens, shows how a relatively long token of /ɪ/, despite a preference for perceiving /long/ rather than /short/, will nevertheless be perceived correctly as /mid, short/.

3.3 Low-level perceptual integration

Once categories have arbitrary labels, the child can consider the relations of each category with *all* auditory continua, not just with one. Thus, there is nothing against including perverse-sounding constraints like "[an F1 of 430 Hz] is not /short/" and "[a duration of 91 ms] is not /mid/", initially low-ranked. Such a procedure is needed for e.g. the integration of [vowel duration], [burst strength], and [closure duration] in the perception of the English word-final obstruent voicing contrast [6].

4. CONCLUSION

By expressing the insights of cognitive-psychological speech perception research with the decision mechanism of the linguistic framework of Stochastic Optimality Theory, our model provides an explicit explanation of the auditory-driven and lexicon-driven mechanisms that underlie the acquisition of language-specific sound categorization.

REFERENCES

[1] Paul Boersma and Bruce Hayes, "Empirical tests of the gradual learning algorithm." *Linguistic Inquiry* 32, 45–86, 2000.

[2] Jessica Maye, Janet F. Werker and LouAnn Gerken, "Infant sensitivity to distributional information can affect phonetic discrimination." *Cognition* 82, B101–B111, 2002.

[3] Kay Behnke, *The acquisition of phonetic categories in young infants: a self-organising artificial neural network approach*. Doctoral thesis, Universiteit Twente [Max Planck Institute Series in Psycholinguistics 5], 1998.

[4] Susan Nittrouer, "Discriminability and perceptual weighting of some acoustic cues to speech perception by 3-year-olds." *JSHR* 39, 278–297, 1996.

[5] Judith E. Pegg and Janet F. Werker, "Adult and infant perception of two English phones", *JASA* 102, 3742–3753, 1997.

[6] Paola Escudero and Paul Boersma, "Modelling the perceptual development of phonological contrasts with Optimality Theory and the Gradual Learning Algorithm." *Proceedings of the 25th Penn Linguistics Colloquium*, to appear.

[7] Paul Boersma, *Functional Phonology*. Doctoral thesis, University of Amsterdam. The Hague: Holland Academic Graphics, 1998.

[8] Diane Kewley-Port, "Thresholds for formant-frequency discrimination of vowels in consonantal context." *JASA* 97, 3139–3146, 1995.

[9] Patricia K. Kuhl, "Human adults and human infants show a "perceptual magnetic effect" for the prototypes of speech categories, monkeys do not." *Perception and Psychophysics* 50, 93–107, 1991.

[10] Frank H. Guenther and Marin N. Gjaja, "The perceptual magnet effect as an emergent property of neural map formation." *JASA* 100, 1111-1121, 1996.

[11] James L. Morgan and Jenny R. Saffran, "Emerging integration of sequential and suprasegmental information in preverbal speech segmentation." *Child Development* 66, 911–936, 1995.

[12] Barbara A. Younger, "Parsing objects into categories: Infants' perception and use of correlated attributes." In D.H. Rakison and L. Oakes (eds.), *Early concept and category development*, ch. 4. Oxford University Press, 2003.

[500 Hz, 104 ms] ⎪ mid, short ⎪	*/high, short/	*/mid, long/	[500 Hz] not /high/	[104 ms] not /short/	[104 ms] not /long/	*/high, long/	*/mid, short/	[500 Hz] not /mid/
/high, long/			*!		*	*		
/high, short/	*!		*	*				
/mid, long/		*!			*			*
√ ☞ /mid, short/				*			*	*