

The evolution of auditory contrast

Paul Boersma
Universiteit van Amsterdam
paul.boersma@uva.nl

Silke Hamann
Utrecht University
silke.hamann@let.uu.nl

April 17, 2007

Abstract. This paper reconciles the standpoint that language users do not aim at improving their sound systems with the observation that languages seem to improve their sound systems. Computer simulations of inventories of sibilants show that Optimality-Theoretic learners who optimize their perception grammars automatically introduce a so-called *prototype effect*, i.e. the phenomenon that the learner's preferred auditory realization of a certain phonological category is more peripheral than the average auditory realization of this category in her language environment. In production, however, this prototype effect is counteracted by an *articulatory effect* that limits the auditory form to something that is not too difficult to pronounce. If the prototype effect and the articulatory effect are of a different size, the learner must end up with an auditorily different sound system from that of her language environment. The computer simulations show that, independently of the initial auditory sound system, a stable equilibrium is reached within a small number of generations. In this stable state, the dispersion of the sibilants of the language strikes an optimal balance between articulatory ease and auditory contrast. The important point is that this is derived within a model without any goal-oriented elements such as dispersion constraints.

1. Introduction

It has often been observed that sound systems are structured in a way that minimizes the perceptual confusion between its elements. For instance, a language with three vowels tends to pronounce them approximately as [a], [i], and [u], three sounds that are maximally far removed from each other in the two-dimensional space of auditory first and second formant. A related observation is the existence of chain shifts in sound change. For instance, a change of a vowel pronounced [u] into a vowel pronounced [y] is often followed by a change of [o] into [u]; or the consonants pronounced [dʰ, d, t] in Proto-Indo-European have shifted to [d, t, θ] (respectively) in Germanic and on to [t, ts, d] (respectively) in German. In all these shifts, the common theme is that the contrast between the members of the inventory is maintained, although all members change. At the abstract level of the language, therefore, it appears that languages actively strive to implement an *optimal auditory dispersion* and to maintain this dispersion diachronically.

At the same time, many authors insist that perceptually-based sound change can only take place by *innocent misapprehension*, i.e. that speakers do not have the perceptual optimization of a sound system as a goal but that sound change is instead caused by learners who reanalyse the imperfectly transmitted sounds of their language environment (Ohala 1981, Blevins 2004).

There seems to be a tension between the idea of optimal auditory dispersion and the idea of innocent misapprehension. A proponent of auditory dispersion even claims that "Sound change through misperception [...] can only hope to account for neutralization, not dispersion or enhancement" (Flemming 2005: 173). This is not entirely true: there may exist non-goal-oriented mechanisms by which improvement of auditory contrast could be a common but unintended *result* of innocent misapprehension. For instance, Blevins (2004: 285–289) tentatively explains chain shifts with the help of Pierrehumbert's (2001) claim that exemplar theory predicts automatic shifting of auditory vowel prototypes to regions where they are less likely to be perceived as a different category (see §7.2). All authors agree, however, that

formalizing auditory dispersion with existing formal phonological devices such as Optimality Theory is incompatible with the claim of innocent misapprehension. Thus, the OT accounts of dispersion by Flemming (1995, 2004), Padgett (2001, 2003ab, 2004), and Sanders (2003) contain explicitly goal-oriented elements, namely dispersion constraints.

The present paper reconciles auditory dispersion with innocent misapprehension in OT. By using a simple bidirectional model of phonetics in which a language user applies the same constraint ranking to perception and production, we show that the innocent misapprehension standpoint can be correct in stating that speakers are not goal-oriented, while at the same time the auditory dispersion standpoint can be correct in observing that sound change tends to minimize perceptual confusion. The reconciliation will be seen to derive from the possibility that sound change is teleological at the abstract level of the observed language but non-teleological at the concrete level of the language user; this situation is analogous to that in evolutionary biology (Darwin 1859), where adaptations to the environment are observationally optimizing but underlyingly non-teleological.

2. Auditory dispersion effects and explanations

In this paper we confine ourselves to the simplest case of dispersion, namely the one-dimensional case. In this section we first present six kinds of dispersion effects, then discuss earlier accounts from the non-OT and OT literature.

2.1 Six kinds of auditory dispersion effects

In Figure 1 we see how phonological categories tend to be dispersed along one-dimensional auditory continua. This figure helps us to illustrate the six kinds of dispersion effects.

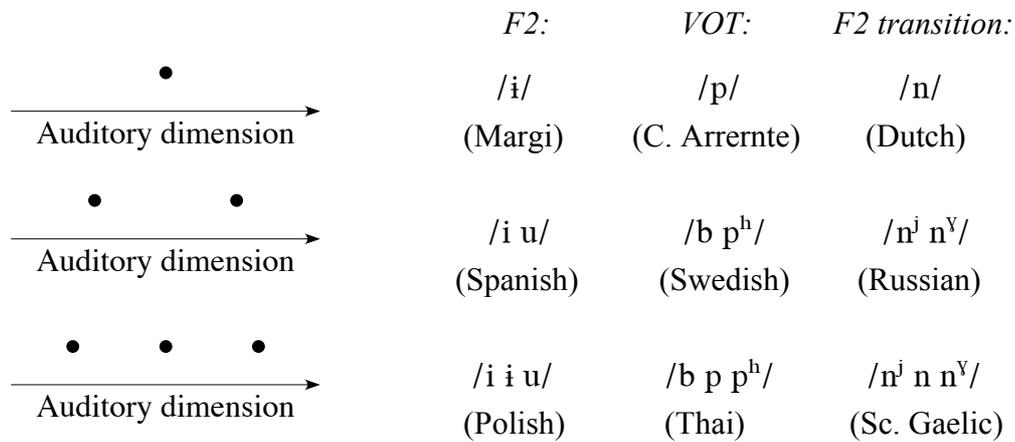


Figure 1

Observed auditory dispersion.

1. The preference for the centre. If a language has only one category on the continuum, it tends to be in the centre (Figure 1, top row). We know this because it tends to have the same auditory value as the mid value of an inventory with three phonemes on the continuum (Figure 1, bottom row). For instance, languages with only one category on the continuum of voice onset time (VOT) for plosives, such as Central Arrernte (Breen & Dobson 2005), typically have only the zero VOT value, as in the plain voiceless plosive [p], while languages with three categories, such as Thai (Tingsabath & Abramson 1999), typically have a ‘prevoiced’ [b] (negative VOT), [p] (zero VOT), and an aspirated [p^h] (positive VOT).

Likewise, languages with only one high vowel, such as Margi (Maddieson 1987) or Kabardian (Choi 1991), tend to have a vowel with a mid second formant (F2) value, such as [i], while languages with three high vowels, such as Polish (Jassem 2003), tend to have a rounded back vowel [u] (low F2), a central vowel [ɨ] (mid F2), and a spread front vowel [i] (high F2). Finally, languages with only one value on the palatalized-velarized continuum for alveolar nasals, such as English, tend to have the plain [n] (mid F2 transition), while languages with three such nasals, such as Scottish Gaelic (Borgström 1940), tend to have a palatalized [nʲ] (high F2 transition), a plain [n], and a velarized [nʷ] (low F2 transition).¹ In all these cases, the single value of the inventories with one category equals the mid value of the inventories with three categories. Two possible explanations have been proposed for such cases. The explanation of *featural markedness* (Jakobson 1941) claims that the mid value ([p] or [n]) reflects a phoneme with *unmarked* feature values (/p/ = [-voiced, -spread glottis], /n/ = [-front, -back]), whereas the peripheral values reflect *marked* feature values (/b/ = [+voiced], /p^h/ = [+spread glottis], /nʲ/ = [+front], /nʷ/ = [+back]). And the explanation of *articulatory effort* (Lindblom 1990) claims that the mid value tends to be the easiest to pronounce ([p] = no active laryngeal gestures, [n] = no special tongue body movements), whereas peripheral values often involve additional gestures ([b] = active vocal fold adduction, [p^h] = active vocal fold abduction, [nʲ] = tongue body fronting, [nʷ] = tongue body backing). The choice between the two explanations has to involve taking into account inventories with two categories on the continuum.

2. The excluded centre. Languages with two categories on the continuum often have them on two sides of the centre. With Flemming (1995 [2002: 37]) and Padgett (2001, 2003ab), we regard this phenomenon as the crucial piece of evidence for the existence of auditory dispersion as a driving force for inventories. Flemming (1995, 2004, 2005, 2006), for instance, notes that a language with two high vowels tends to have [i] (high F2) and [u] (low F2), crucially excluding the intermediate [ɨ] vowel, which is the one that languages with one high vowel such as Margi and Kabardian tend to have. He then argues that the usual account of inventories in terms of markedness of phonemes does not work: since [i] and [u] are much more common cross-linguistically than [ɨ] is, and [i] slightly more common than [u], a markedness account would probably regard /i/ as the unmarked high vowel and would therefore predict that languages with a single high vowel have /i/. Padgett makes a similar argument for the palatalization-velarization continuum in Russian, which has [nʲ] and [nʷ] but not the arguably unmarked and least effortful [n]. The only explanation is that in the [i]-[u] pair and in the [nʲ] and [nʷ] pair the concept of auditory contrast plays the decisive role. With featural markedness out of the game, the explanation of the [p], [i] and [n] singletons discussed in the previous paragraph must be articulatory.

3. Equal auditory distances. The two effects just mentioned can be summarized as a single *primary auditory dispersion effect*: categories tend to be located within the auditory space in such a way that they are perceptually maximally distinct. For languages with three categories along a continuum, this idea predicts that they should typically have the middle category spaced at equal auditory distances from its neighbours. We can check this when looking at languages with four categories along the continuum. For instance, an optimal dispersion of the auditory vowel height continuum (first formant, F1) involves that the middle

¹ In Bernera Gaelic, Ladefoged, Ladefoged, Turk, Hind & Skilton (1997) did not find the three-way contrast on the nasals, but they did find it on the laterals.

value of the triplet [i “e” a] (e.g. Bradlow 1995 for Spanish) lies in between the two mid values of the quadruplet [i e ε a] (e.g. Harrison 1997 for Catalan).

4. The growing space. A *secondary auditory dispersion effect* is that for larger inventories, the auditory space enlarges, but the distance between the categories decreases. In Figure 1, for instance, the auditory space taken up by the inventory with three elements is larger than the auditory space that has to accommodate only two elements. For instance, the high back vowel is often auditorily fronted in languages with two high vowels (e.g. English [ɨ:], Japanese [ɯ]), but not in languages with three high vowels. In languages with many vowels, the three corners of the vowel space tend to be [a, i, u] whereas languages with only three vowels often have the reduced [ɐ, ɪ, ʊ] (Boersma 1998: 216); and languages with no more than a single vowel (such as several Germanic languages in unstressed syllables) typically have just [ə] (Flemming 2004: 235, 2005: 164). The explanation is that the more peripheral a value is, the more articulatory effort tends to be required to implement it ([ə] corresponds to a neutral tongue shape); languages with two categories require a smaller total auditory space for obtaining a small perceptual confusability than languages with three categories do, so that the balance between auditory contrast and articulatory effort turns out differently for the two cases (note that this is a goal-oriented description at the level of the language, not implying that there is any goal-orientedness in the underlying mechanism).

5. Allowed variation. If the phenomena described in the previous paragraphs really have to do with auditory contrast, this would predict that a category is allowed more auditory variation if it is alone on its auditory continuum than if it has neighbours from which it has to stay distinct. This is borne out by the data. Languages with a single labial plosive [p] typically have voiced allophones such as [b] in post-nasal or intervocalic position (for Central Arrernte: Breen & Dobson 2005), and languages with a single high vowel [i] typically have allophones everywhere between [i] and [u], depending on the surrounding consonants (for Kabardian: Choi 1991). There is no doubt that a syntagmatic articulatory optimization is involved. Again we see an interplay between the demands of auditory contrast and the demands of articulatory economy.

6. Chain shifts. The five effects just mentioned are all *static* effects: they can be seen in synchronic inventories. There also exist two kinds of *dynamic* effects, which can be seen in the diachronic development of inventories: in a *push chain*, one category approaches another and seemingly pushes it away, and in a *drag chain*, one category vacates a region on the auditory continuum, thereby seemingly allowing another category to fill up the vacated space. In order to account for these chain shifts, a model has to exhibit properties that lead to a repulsive force between categories. Not all explanations of inventories have this property. In some theories of inventories a restricted kind of dispersion comes about by selective neutralization of categories (categories that are auditorily close to each other have a greater chance of merging into a single category, so that the remaining categories tend to be spaced further apart) or by unsupervised *clustering* of proto-categories (e.g. De Boer 1999, Oudeyer 2006). Such theories, in which close categories attract rather than repel one another, can only account for diachronic merger, not for chain shifts, and must therefore be left out of consideration in the present paper, because even the strongest proponents of innocent misapprehension agree that “there is no question that chain shifts exist” (Blevins 2004: 285).

It must be clearly stated here that all these effects are tendencies rather than fixed rules. For one thing, the language may be in a transitory, non-equilibrium state, and this may involve non-optimal auditory dispersion (as will become clear in our simulations in §5 and §6). Next, considerations of articulatory effort may limit auditory contrast to the extent that an

inventory with two categories can include the centre value, so that it is asymmetric along the auditory continuum. For instance, while the optimally dispersed [b p^h] inventory does exist in Swedish (Jakobson 1941 [1962: 329], Ringen & Helgason 2004), it does involve two separate articulatory gestures and may therefore be replaced with [b p] (Dutch, French), [b̥ p^h] (English, German), or [p p^h] (Mandarin, Scottish Gaelic), all of which are articulatorily easier and still have a sufficient auditory contrast (Lindblom 1990 calls this effect *adaptive dispersion*). Finally, these single auditory continua live in an inventory with multiple continua, some of which may intrude. For instance, Flemming (1995 [2002: 31]) explains the fact that in many languages /i u/ is realized as [i ʊ] or [i ɯ] as an enhancement of the /u/-/o/ distinction. Rather than contradicting the auditory dispersion idea, all these exceptions corroborate it by leading to explanations that involve articulatory effort, wider-scoped data, and contingent histories, beside auditory dispersion, but never featural markedness.

The following sections discuss various ways in which auditory dispersion has been modelled explicitly.

2.2 Non-OT accounts of auditory dispersion

Auditory dispersion in vowel inventories was first modelled explicitly by Liljencrants & Lindblom (1972), who showed that two-dimensional vowel spaces (auditory height and backness) with a minimum probability of perceptual confusion look like real attested vowel inventories (some problems of detail were addressed by Lindblom 1986 and Schwartz, Boë, Vallée & Abry 1997).

Ten Bosch (1991) compared several techniques for optimizing the distances between vowels, and found that the strategy of *maximizing the minimal distance* worked best. If one starts out with a random set of vowels and then iteratively moves apart the two vowels that are closest to each other, one ultimately obtains an evenly dispersed inventory.

A possible criticism of this work is that the resulting vowel inventories are not symmetric enough. For instance, real vowel inventories tend to be structured in such a way that back vowels tend to have the same height contour as front vowels, yet there is no force in these models according to which such symmetries are enforced, and indeed the vowel inventories that result from the simulations of these authors tend to be asymmetric (Boersma 1998: 357). To remedy this problem, the simulations would have to be extended with tricks to incorporate an efficient use of available features.

A problem that is much more difficult to remedy is the inherent teleology in these models. In order to arrive at an optimal vowel inventory, a model typically starts with a random non-optimal vowel inventory and tries to move towards an optimal end result by making small changes to the locations of the vowels in the vowel space, where every change has to be optimizing, i.e. every change has to improve the auditory distinctivity of the whole system. Even at the most concrete level of modelling, then, these models work with teleological devices.

Non-teleological accounts of auditory dispersion have been proposed as well. Blevins (2004: 285–289) sketches how within a framework where listeners store auditory events as exemplars in episodic memory and subsequently reuse these exemplars in production (Pierrehumbert 2001) speakers may tend to choose exemplars that are little likely to be perceived as anything but the intended category. An explicit account of the details of such a scheme is provided by Wedel (2004, 2006). See also §5.5 and §7.2.

Few of the non-OT accounts mentioned in this section make contact with phonological theory. No interaction of the simulated inventories with phonological rules or constraints is

modelled or even predicted, although the exemplar models might in the future provide such a link. In the next section we discuss attempts to integrate the dispersion idea into a tested and tried framework for phonological theory, namely Optimality Theory.

2.3 OT accounts of auditory dispersion

The original proposal of Optimality Theory by Prince & Smolensky (1993) handled inventories by the device of *Richness of the Base* and constraint interaction. As in the featural markedness accounts mentioned above, a marked phonological element was only allowed to surface in a language if the unmarked counterpart of that phonological element also surfaced in that language. Prince & Smolensky's approach is therefore problematic for the same reason as mentioned earlier: it cannot account for the "excluded centre" effect, where marked segments appear without the unmarked counterpart, as noted by Flemming (1995 [2002: 37], 2004: 235, 2005: 164, 2006: 250). So an OT account requires something more than just markedness and faithfulness constraints.

Flemming (1995) translated Lindblom's dispersion idea, and specifically Ten Bosch's idea of maximizing the minimum distance, into OT by introducing MINDIST constraints, which explicitly militate against inventories with small auditory distances between its members. Dispersion constraints such as Flemming's MINDIST, as well as the reformulations by Padgett (2001) and Sanders (2003), work very well in formalizing the dispersion idea. However, they are explicitly teleological with respect to auditory dispersion. Furthermore, an empirical problem is that these constraints evaluate multiple inputs at a time: they can be said to evaluate whole inventories (Flemming 1995 [2002: 33–35], Boersma 1998: 361, McCarthy 2002: 226–227) or even entire languages (Padgett 2003a: 311, Flemming 2004: 268). These constraints are therefore hard to reconcile with the single-input constraints introduced by Prince & Smolensky, and the tableaux are hard to reconcile with tableaux that basically evaluate the processing of a single form in production (Prince & Smolensky 1993) or comprehension (Smolensky 1996). The general defence by Flemming and Padgett is that phonological theory is about possible languages, rather than about processing single forms. We feel that this standpoint underestimates the power of Optimality Theory as a decision mechanism: when used to evaluate single inputs, OT can be and has been applied successfully to processes such as production, comprehension, and acquisition. If dispersion effects can be shown to emerge in OT even from modelling single-form processing, OT will not have to be invoked separately to evaluate the entire language.

The fact that Flemming, Padgett and Sanders handle dispersion effects by dedicated inherently teleological means (the dispersion constraints) in a synchronic grammar is sometimes seen as unproblematic (e.g. Hayes & Steriade 2004: 27), but a theory in which these effects arise automatically as side effects of more general independently needed devices should be preferred by Occam's razor if such a theory exists. Padgett (2003b: 80) realizes this shortcoming and suggests that dispersion constraints may just express abstract observations about inventories while at the same time the real underlying mechanism may be more concrete and perhaps not explicitly goal-oriented. We agree, and in the present paper we provide just such an underlying mechanism. Providing the underlying mechanism is necessary because we seek the locus of explanation in an acquisition bias on the part of the learner; formalizing acquisition has to be done with the constraints that are in the learner's brain, not with constraints that describe behaviour at a higher level of abstraction.

In the present paper, we employ neither dispersion nor faithfulness constraints to account for dispersion effects. We show that dispersion effects can instead arise within a number of

generations as the automatic result of *cue constraints*, which are independently needed to model language-specific perception (e.g. Escudero & Boersma 2004), and *articulatory constraints*, which are independently needed to model articulatory effort in phonetic implementation (e.g. Kirchner 1998, Boersma 1998). The only assumption that we need to add to the pre-existing work on OT phonetics is that the speaker and the listener use the same grammar. The point is that the same constraints are used *bidirectionally*, i.e. both by the listener in comprehension and by the speaker in production. The following section illustrates this in more detail.

3. Bidirectional phonetics

A formal account of a linguistic phenomenon has to start by stating the representations involved. For our purposes we need only two, namely the (abstract, discrete) phonological *surface form* and a (concrete, continuous) auditory-articulatory *phonetic form*. These two representations (see Fig. 2) are part of a more elaborate comprehensive model for bidirectional phonology and phonetics (Boersma 2005), but any representations ‘above’ the surface form, such as the underlying form, are not discussed in this paper (except briefly in §5.3 and §7.1) because they are not required for illustrating our point. Also, Figure 2 keeps implicit any distinction between the auditory part and the articulatory part of the phonetic form.

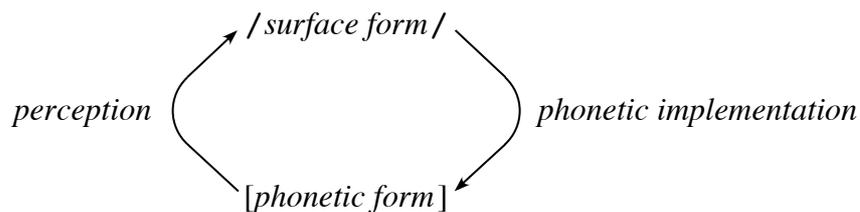


Figure 2

Processing models for phonetics and its interface with phonology.

We propose that the grammar model of Figure 2 is bidirectional, i.e. that it is used in two directions of processing: comprehension and production, according to the arrows in the figure. In the comprehension direction, the (‘prelexical’) *perception* process maps an auditory-phonetic form to a phonological surface form; in the production direction, the *phonetic implementation* process maps a phonological surface structure to an auditory-articulatory phonetic form.

We formalize the representations in Figure 2 as well as the relations between them within the framework of Optimality Theory, as in Figure 3.

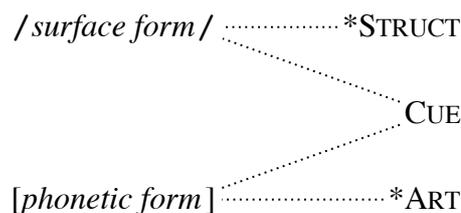


Figure 3

Grammar model for phonetics.

One of the reasons for using OT to model phonetics is that the output of the perception process tends to be constrained by the same structural constraints (*STRUCT in the figure) that have been proposed for phonological production. For instance, Polivanov (1931) proposes that Japanese learners of Russian perceive the Russian phonetic form [tak] as /.ta.ku./ because of a Japanese constraint on coda consonants, and the phonetic form [drama] as /.do.ra.ma./ because of a Japanese constraint against complex onsets.² Thus, the process that maps a phonetic form (or ‘overt form’ in the terms of Tesar & Smolensky 2000) to a phonological surface form is best regarded as being linguistic itself and therefore amenable to OT modelling, a point made before by Tesar & Smolensky (1998, 2000), Boersma (1998), and Pater (2004).

We like to point out here that modelling phonetic processing within Optimality Theory does not imply that we regard low-level auditory and articulatory processing as belonging to a formal symbolic system specific to the human language faculty. On the contrary, a constraint-based decision mechanism like OT may well be typical of neural processing in general, and the success of OT in phonological theory may well be based on the fact that human phonological processing just uses this more general mechanism.

Having made plausible that phonetic processing can be modelled with Optimality Theory, we can turn to the constraints proposed in Figure 3. The relation between the phonological and the phonetic form is evaluated by cue constraints (*CUE), and the phonetic form on its own is evaluated by articulatory constraints (*ART). As an example of a cue constraint, consider the fact that the duration of a vowel is a major cue to the voicing of a following obstruent in English but not in most other languages, both in perception (Denes 1955, Raphael 1972) and in production (Heffner 1937, House & Fairbanks 1953). Hence, the cue constraint *[long vowel duration]/obs,-voice/ is ranked high in English but low elsewhere.

We propose that the constraints in Figure 3 are used bidirectionally, i.e. that they are used both in perception and in phonetic implementation. Bidirectionality in OT was proposed earlier for faithfulness constraints by Smolensky (1996), for structural constraints by Tesar (1997), Tesar & Smolensky (1998, 2000) and Pater (2004), and for cue constraints by Boersma (1998).

In perception, the choice between candidate surface forms involves structural and cue constraints. In this direction of processing, the cue constraints evaluate language-specific cue integration (Escudero & Boersma 2004). For instance, the high ranking of *[long vowel duration]/obs,-voice/ in English predicts that an English listener, when confronted with an auditorily lengthened vowel, will be unlikely to perceive the following consonant as a voiceless obstruent, unless the possibly conflicting constraints for other cues (or perhaps competing structural constraints) force her to.

In phonetic implementation, the choice between candidate phonetic forms involves cue constraints and articulatory constraints. For instance, the high ranking of *[long vowel duration]/obs,-voice/ in English predicts that an English speaker, when intending to realize a voiceless obstruent, will be unlikely to lengthen the preceding vowel, unless articulatory constraints force her to.

We thus propose that cue constraints are used both in perception and in phonetic implementation, and that these constraints are ranked identically in both directions of

² Polivanov’s proposal was confirmed in perception experiments by Dupoux *et al* (1999) and in brain activity experiments by Jacquemot *et al* (2003). A reformulation in OT terms appeared in Escudero & Boersma (2004).

processing. The present paper shows that this bidirectional use of cue constraints leads to two asymmetries between perception and production, namely the *prototype effect* and the *articulatory effect*, and that languages that are stable over the generations have to cancel these two biases out against one another, thus striking an optimal balance between minimization of articulatory effort and minimization of perceptual confusion, without there being any goal-oriented dispersion mechanism in the whole system.

4. Sibilant inventories: dispersion of the spectral mean

The sounds with which we will exemplify the evolution of auditory contrast in this paper are the sibilants. We first make it plausible that sibilants indeed form a case of auditory dispersion.

The sibilants in a language can often be ordered along a one-dimensional auditory continuum, namely the *spectral centre of gravity* or *spectral mean* (e.g. Forrest, Weismer, Milenkovic & Dougall 1988; Gordon, Barthmaier & Sands 2002).³ Articulatorily, the spectral mean correlates with the frontness of the tongue and with the frontness of the place of articulation.

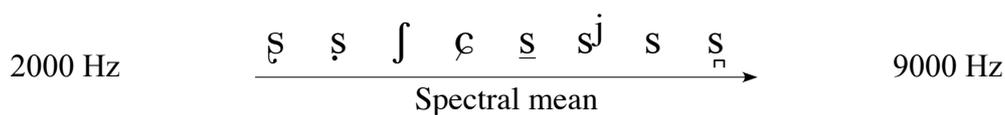


Figure 4

Scale of spectral mean for eight sibilants.

We pretend in this paper that spectral mean is the only relevant perceptual cue for sibilants. In reality, other cues for sibilant articulator and place include spectral peaks (e.g. Jongman, Wayland & Wong 2000) and vowel transitions (e.g. Nowak 2006).

Sibilant inventories tend to be dispersed along this dimension, exhibiting all the effects mentioned in §2.1. Languages with only one sibilant, such as Spanish or Dutch (if we disregard the marginal and unstable Dutch alveolopalatal, but see fn. 10) usually employ a sound with a fairly central spectral mean such as the Dutch flat laminal alveolar [s̺] (Mees & Collins 1982: 6) or the Spanish concave retracted apical alveolar [s̺̄] (Navarro Tomás 1932: 105–107, Harris 1969: 192). This is illustrated at the top of Figure 5. If a language has two sibilants, neither of them has a central spectral mean. Both sibilants are rather peripheral on the spectral-mean dimension, such as in English, which has a laminal, shallow-grooved and often rounded postalveolar [ʃ] and a deep-grooved alveolar [s] (Stone, Faber, Raphael & Shawker 1992: 260).

³ Values reported in the literature can vary widely, which is partly due to the measurement method involved. In this paper, we assume that the spectral mean is computed by weighing the frequencies in the spectrum by their power densities (Forrest *et al* 1988, Jongman *et al* 2000). This is the standard setting in the Praat program (Boersma & Weenink 1992–2007), from which it was used by Zygis & Hamann (2003) and in the OT modelling by Padgett & Zygis (2003). The method yields values for sibilants that can indeed be as low as 2000 Hz or as high as 8000 Hz or more. As one can judge from their spectra and reported spectral means, Gordon *et al* (2002) apparently used the incorrect method by Ladefoged (2003), which weighs the frequencies by their intensity values in dB and is therefore sensitive to arbitrary recording settings; this method tends to yield values very close to the centre of the frequency range, i.e. in Gordon *et al*'s case very close to 5000 Hz.

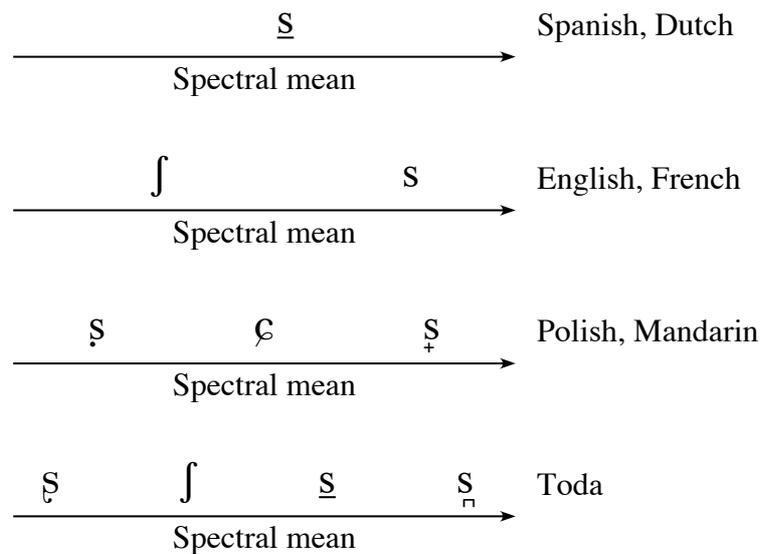


Figure 5

Dispersion of the spectral mean for inventories with one, two, three and four sibilants.

In languages with three sibilant categories, these sounds are equally spaced along the spectral-mean dimension. At the same time, these three sibilants occupy a larger space than a two-sibilant inventory, illustrating the secondary dispersion effect (compare the third with the second graphic in Figure 5). This holds for instance for the sibilant inventories of Polish (Puppel, Nawrocka-Fisiak & Krassowska 1977) or Mandarin (Ladefoged & Wu 1984) with (denti-)alveolar, alveopalatal and apical postalveolar sibilants (the Polish inventory was explicitly described in terms of auditory dispersion by Jones 2001, Zygis 2003, and Padgett & Zygis 2003). Similarly, languages with four sibilants such as Toda with a dental, an apical alveolar, a laminal postalveolar and a subapical palatal sibilant (Shalev, Ladefoged & Bhaskararao 1993), occupy an even larger space along the dimension of spectral mean. The distances between adjacent sibilants are equally large, but they are smaller than those in inventories with three sibilants. Presumably, producing extremely low or extremely high spectral-mean values involves more articulatory effort than producing central values.

In the following two sections, we show how the dispersion of sibilant inventories can be modelled without dispersion constraints. We illustrate this for English, a language with two sibilants, and for Polish, a language with three sibilants that show chain-shift effects, in sections 5 and 6, respectively.

5. The English two-sibilant inventory

English has two sibilants, an alveolar /s/ and a postalveolar /ʃ/. In this section we first show how the English auditory environment leads us to predict the properties of an optimal English OT listener. We describe in detail how an English learner comes to be an optimal listener of her language, and perform a computer simulation showing that a virtual English learner arrives at a pronunciation that matches that of her environment. Having thus shown that English is a stable language, we next show that a language with the exaggerated sibilant inventory /ʃ/-/ʃ/ is not stable but will instead turn into English within three generations. We do the same for the skewed and confusable inventory /ʃ/-/s/.

5.1 The English auditory language environment

For simplicity we assume here that the auditory difference between the two English sibilants is wholly caused by a difference in their spectral means. The spectral mean of e.g. /s/ will vary between speakers, vowel environments, and acoustic and auditory conditions, as well as between replications by the same speaker. A listener will be able to normalize away some of this variation, but not all of it. Since the non-normalizable variation has multiple sources, the distribution of normalized spectral means for all the /s/ tokens that a listener hears is likely to have a bell-like shape, perhaps similar to the Gaussian shape in Figure 6. In order to be able to work with round numbers in our simulations, we fix the average spectral mean of all normalized /s/ tokens in the listener's language environment at a slightly arbitrary value of 7000 Hz. Likewise, we assume for /ʃ/ a similar Gaussian distribution with an average spectral mean of 4000 Hz.⁴

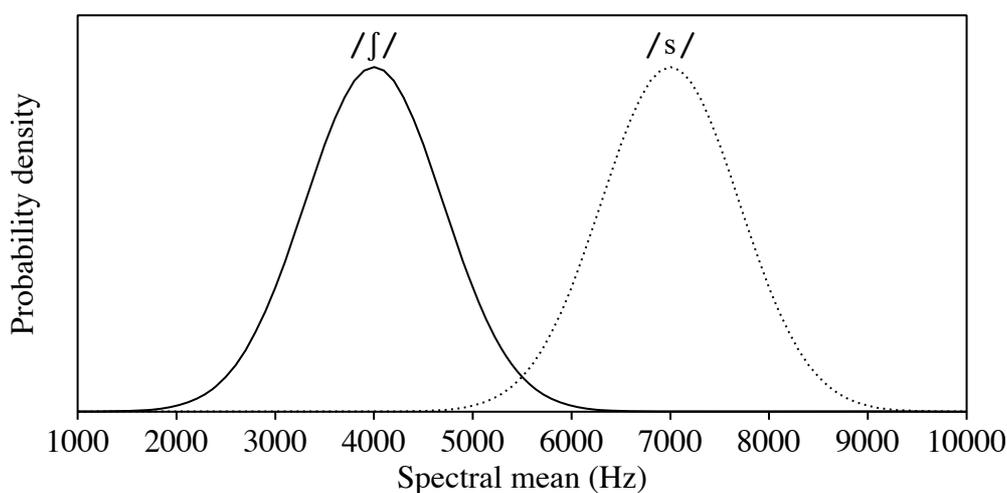


Figure 6

Distributions for tokens of the English sibilants /ʃ/ and /s/ along the spectral mean.

The two distributions in Figure 6, then, are the material that an adult listener has to work with, and that an infant learner has to make sense of.

5.2 The adult as an optimal listener

We explicitly assume that adult English listeners have learned to minimize the probability of misunderstanding, i.e. that they have become *optimal listeners*. To minimize the probability of misunderstanding, the optimal listener must show *maximum-likelihood behaviour* (Fisher 1922) i.e. she must classify an auditory event as the category that is most likely to have been intended by the speaker. For sibilants, the optimal English listener's decision criterion must be the frequency where the two curves in Figure 6 intersect: she must perceive all tokens with a spectral mean below [5500 Hz] as /ʃ/, and all tokens above [5500 Hz] as /s/.

⁴ Jongman *et al* (2000) report the values of 6133 and 4229 Hz, respectively (for causes described in footnote 3, Gordon *et al* 2002 typically report differences of only 400 Hz between /s/ and /ʃ/ for the languages they investigate). We use round numbers instead, because our goal is to make a proof of principle, not to be numerically realistic. Had we wanted to be more realistic, the frequency scale would have had to be in psychoacoustic (nearly logarithmic) units instead of Hertz.

To model this correlation of spectral-mean values and surface categories, we follow Escudero & Boersma (2004) in using a complete set of arbitrary cue constraints, which are summarized in (1).

- (1)
- | | |
|----------------|----------------|
| *[1000 Hz]/s/ | *[1000 Hz]/ʃ/ |
| *[1100 Hz]/s/ | *[1100 Hz]/ʃ/ |
| ... | ... |
| *[9900 Hz]/s/ | *[9900 Hz]/ʃ/ |
| *[10000 Hz]/s/ | *[10000 Hz]/ʃ/ |

For perception, the constraint *[1000 Hz]/s/ can be read as “an auditory spectral mean of [1000 Hz] should not be perceived as the surface phonological category /s/”; for phonetic implementation, the constraint can be read as “a surface form /s/ should not be realized with a spectral mean of [1000 Hz].” For simplification, we discretize the spectral-mean range into 91 steps of 100 Hz. In reality, the resolution is based on hair cells and auditory nerve fibers. Combining these 91 spectral-mean values with the two sibilant categories results in a total of 182 cue constraints.

In connecting every possible spectral-mean frequency between [1000 Hz] and [10000 Hz] to both sibilant categories, the cue constraints in (1) are very different from usual OT constraints, which either tend to express typological trends directly (e.g. markedness constraints) or tend to express universally preferred relations (e.g. faithfulness constraints). That is, the set in (1) has no preference for certain spectral-mean values, for certain sibilants, or for connecting certain spectral mean values to certain sibilants, i.e., the constraints are not restricted to actually occurring spectral-mean values or to actually occurring combinations of spectral means and sibilant categories. It is the *ranking* of these constraints that will have to be responsible for making the correct connections for English. From Figure 6 we see, for example, that an optimal listener of English should perceive [4700 Hz] as /ʃ/. A possible ranking that achieves this perception is given in the perception tableau in (2).

(2) *A perception tableau for classifying tokens with a spectral mean in English*

[4700 Hz]	*[4600]/s/	*[4700]/s/	*[4800]/s/	*[4800]/ʃ/	*[4700]/ʃ/	*[4600]/ʃ/
/s/		*!				
☞ /ʃ/					*	

The correct classification of [4700 Hz] is seen to rely solely on the ranking *[4700]/s/ >> *[4700]/ʃ/. In the same way we can establish rankings for the 90 remaining spectral mean values, e.g. *[7300]/ʃ/ >> *[7300]/s/; the constraint *[5500]/s/ will be ranked at approximately the same height as *[5500]/ʃ/. We have now shown that the 182 constraints can be ranked in such a way that we can model an optimal listener for the distributions of Figure 6.

However, such a language-specific ranking of 182 constraints is rather uninformative when it comes to predicting what are possible kinds of sibilant categorization strategies in the languages of the world. For one thing, the set of 182 constraints seems to generate an incorrect factorial typology: the ranking { *[3100]/s/, *[3200]/ʃ/, *[3300]/s/, *[3400]/ʃ/ } >> { *[3100]/ʃ/, *[3200]/s/, *[3300]/ʃ/, *[3400]/s/ }, for instance, describes a language in

which [3100 Hz] is perceived as /ʃ/, [3200 Hz] as /s/, [3300 Hz] again as /ʃ/, and [3400 Hz] again as /s/. In other words, the constraint set allows categories with massively non-contiguous auditory correlates. Such languages have not been shown to exist, a fact that calls for an explanation. If this explanation cannot be provided by ranking permutation, what else can provide the explanation? The general answer is based on the idea that if UG allows grammars that are not attested in reality, then those grammars might be unlearnable or unstable over the generations. We will therefore model both the acquisition and the evolution of these grammars. The following subsection illustrates how the acquisition of English sibilants is modelled with a simple learning procedure and algorithm.⁵

5.3 Learning to become an optimal listener of English

We describe here the situation when a child already has correct lexical representations, but not yet an adult-like prelexical perception. That is, she already knows which lexical items have an underlying |ʃ| and which have an underlying |s|, but she does not know in the adult-like way of §5.2 what spectral-mean values occur with which of the two surface sibilants.

During this acquisition period, the child will receive many tokens of /ʃ/ and /s/ drawn from the distributions in Figure 6. Some of the time she will make a mistake, as in tableau (3), which occurs when an adult speaker talking to the child pronounces an intended |ʃ| with a reasonable spectral mean of 4700 Hz.

(3) A learner's perception tableau with reranking of cue constraints

[4700 Hz]	*[4600]/s/	*[4800]/ʃ/	*[4700]/ʃ/	*[4600]/ʃ/	*[4700]/s/	*[4800]/s/
☞ /s/					←*	
√ /ʃ/			*!→			

The child of tableau (3) has the non-optimal ranking $*[4700]/ʃ/ \gg *[4700]/s/$ and therefore perceives the incoming auditory event [4700 Hz] as the vowel category /s/, as indicated by the pointing finger (☞). However, the speaker had intended to transmit the lexical symbol |ʃ|, and the semantic and pragmatic context may lead the child's comprehension system to realize this (perhaps because the recognized lexical item was *sheep*). As a result, the child's lexicon can subsequently 'tell' her that she should have perceived /ʃ/ instead of /s/. This new knowledge by the child is indicated by the check mark (√) in the tableau.

When a perception tableau such as (3) contains the child's own winning form (☞) as well as a form that she considers correct (√), and the two forms are different, the child can conclude that she made a mistake. As a result, she can take action by taking a *learning step*. A good strategy for executing a learning step is the *Gradual Learning Algorithm* (Boersma 1997, Boersma & Hayes 2001), which is indicated by the arrows in the tableau: all constraints that are in favour of the correct category are moved up, and all constraints in favour of the child's own 'incorrect' winner are moved down. As a result, the two cue constraints for the value [4700 Hz] will be reranked slightly in the direction of the arrows. This process is called *lexicon-driven perceptual learning* (Boersma 1997, Escudero & Boersma 2004).

⁵ For a simulation on the learnability of non-contiguous categories, see §6.4.

The arrows in tableau (3) represent small steps along a continuous scale of ranking. After having heard a number of [4700 Hz] events that should have been perceived as /ʃ/, the learner will have swapped the rankings of *[4700]/ʃ/ and *[4700]/s/. From that time on, the learner will perceive the auditory form [4700 Hz] correctly as the category /ʃ/. The same successful learning applies to all other auditory forms for which one of the curves in Figure 6 is close to zero, i.e. for all forms below approximately [4700 Hz] or above approximately [6300 Hz]. It now becomes clear why we called the Gradual Learning Algorithm a ‘good strategy’: the learner has become a maximum-likelihood listener for those auditory values. In the region where the curves in Figure 6 overlap, something slightly different happens: the listener will necessarily continue to make some mistakes in prelexical perception, simply because e.g. an incoming token of [5300 Hz] was intended as |ʃ| 75% of the time, but as |s| 25% of the time, as the curves show. In *Stochastic Optimality Theory* (Boersma 1997, Boersma & Hayes 2001), where the ranking of every constraint is subject to a bit of additive noise at evaluation time, the listener is likely to vary her perceptual decisions. The Gradual Learning Algorithm then leads to a situation of *probability matching*: the learner will end up in an equilibrium situation in which she perceives [5300 Hz] as /ʃ/ 75% of the time, but as /s/ 25% of the time. This works because the constraint *[5300]/s/ will end up being ranked just above *[5300]/ʃ/. This strategy of probability matching in perception, which is nearly as good as the maximum-likelihood strategy, has been shown to emerge automatically from the lexicon-driven learning mechanism both for one-dimensional continua (Boersma 1997) and for two-dimensional continua (Escudero & Boersma 2003).

In order to see exactly what happens in perception, and to be able to predict the effect of lexicon-driven perceptual learning on phonetic implementation, we cannot limit ourselves to the present description of the learning mechanism. Instead, a computer simulation is required.

5.4 Simulation of the acquisition of English perception

The learning of an English perception grammar can be simulated with the computer. Our virtual learner has the 182 cue constraints in (1), and starts out having them all ranked at the same height of 100.0; as a result, the learner has an initial state where she perceives every incoming spectral-mean value as /s/ 50% of the time and as /ʃ/ 50% of the time. At the same time we assume that this initial virtual learner already has correct lexical representations for all words with /s/ and /ʃ/. Although this combination of fully random prelexical perception and perfect lexical storage is obviously unrealistic, a more realistic modelling is very likely to exhibit effects that work in the same directions as the ones we will find here, although their size may differ.⁶

During the simulated acquisition period, our learner hears 1 million /ʃ/ and 1 million /s/ tokens in random order and with spectral-mean values that are randomly drawn from the distributions in Figure 6. Each token is also labelled as /s/ or /ʃ/ by the learner’s lexicon. When hearing a token, the learner will perceive it into either category. In our simulations we set the standard deviation of Stochastic OT’s *evaluation noise* (per-tableau random variation in ranking) to a constant value of 2.0.

It can happen that the learner’s perceived category is identical to the category the lexicon tells her she should have perceived. In such a case, the learner does not change her grammar. But if the perceived category is different from the one her lexicon says is correct, our learner

⁶ See §7.3 for a discussion.

changes the ranking of her cue constraints according to the scheme in tableau (3). The Gradual Learning Algorithm is here taken to have a constant *plasticity* (reranking step) of 0.01.

After listening to the 2 million tokens and some considerable shuffling of cue constraints, our learner ends up with the perception grammar in Figure 7. In this figure, the constraints for the sibilant /ʃ/ are connected with a solid curve, and those for /s/ are connected with a dotted curve. We can see, for instance, that for a spectral mean value of [3000 Hz], the dotted curve has a ranking value of 103.5, and the solid curve one of 96.5. This means that the cue constraint *[3000 Hz]/s/ is ranked much higher than *[3000 Hz]/ʃ/, and thus the token [3000 Hz] is most often perceived as /ʃ/ in this perception grammar (not always, because of the evaluation noise).

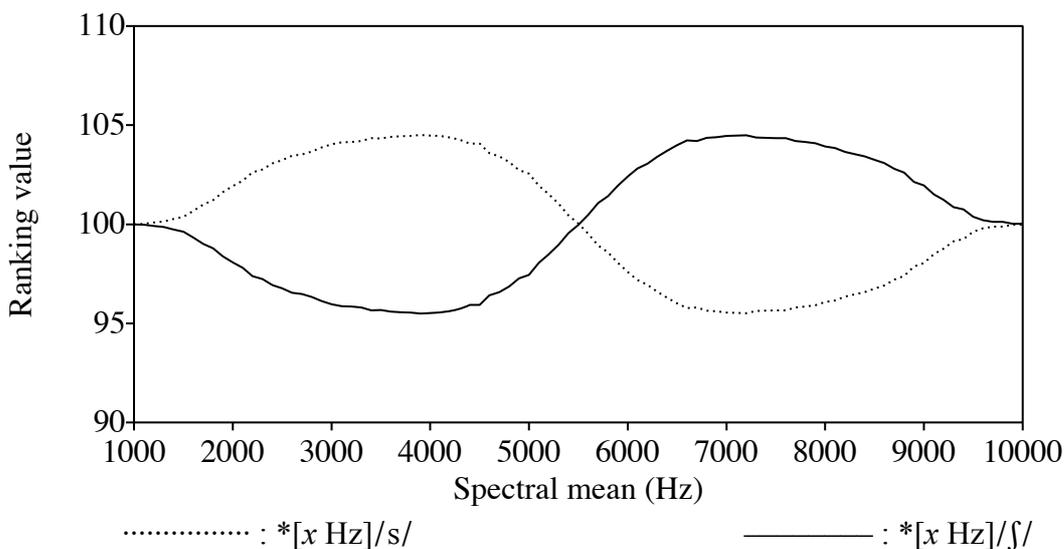


Figure 7

The virtual learner's final perception grammar for the two sibilant categories of English.

In general, the lowest curve determines which category is perceived most often at any specific spectral-mean value in Figure 7.

A complete OT analysis of a problem generally involves two parts, namely a description of the constraint ranking and a description of what outputs the grammar assigns to all of its inputs. While the ranking of all 182 constraints is given in Figure 7, a description of the workings of the grammar would involve giving 91 input-output pairs, perhaps in the form of 91 perception tableaux that are analogous to tableau (2). And even these 91 tableaux would not suffice, because as a result of the evaluation noise of Stochastic OT it is fully possible that the same spectral-mean value is sometimes perceived as /ʃ/, sometimes as /s/. A full account of how the 91 spectral-mean values are handled by the perception grammar therefore involves giving for each of the 91 spectral-mean values the probability that it is perceived as /ʃ/ and the probability that it is perceived as /s/. An estimate of these is given in Figure 8. This figure has been computed by running each of the 91 spectral-mean values through the perception grammar 100,000 times (with an evaluation noise of 2.0, as during learning), and noting how often the output was /ʃ/; dividing each of the 91 results by 100,000 yields an estimate of the

probability that each spectral-mean is classified as /ʃ/.⁷ The probability of classifying it as /s/ is estimated in an analogous way.

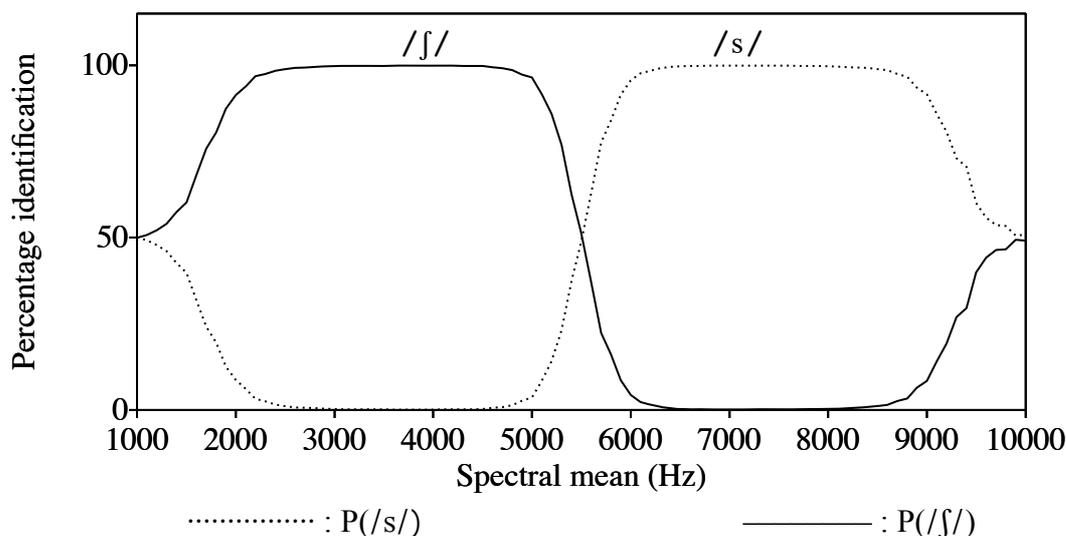


Figure 8

Identification curves for the virtual listener of Figure 7.

In Figure 8 we see that the probability of perceiving /ʃ/ is greatest for those spectral values x for which the curve of $*[x \text{ Hz}]/ʃ/$ lies farthest below the curve $*[x \text{ Hz}]/s/$ in Figure 7. This is a general property of Stochastic OT: the farther constraint A is ranked above constraint B, the smaller the chance that B will overcome A at evaluation time.

Figure 8 confirms that the Gradual Learning Algorithm leads to optimal perception if the learner is given sufficient information. In the regions where she has heard a large number of spectral-mean tokens (i.e. between approximately 2500 and 8500 Hz, according to Figure 6), Figure 8 shows that the learner has become an optimal (probability-matching) listener. For instance, the spectral mean of 5300 Hz has a 75% probability of being perceived as /ʃ/ and a 25% chance of being perceived as /s/; hence, the odds of perceiving 5300 Hz as /ʃ/ are 3 times higher than the odds of perceiving it as /s/; this ratio of 3 to 1 corresponds exactly to the relative heights of the two curves in Figure 6 at [5300 Hz].

However, the learner does not become optimal if she is given too little information. In the left periphery of the auditory space (around 1500 Hz), the probability that such a spectral-mean value was intended as /ʃ/ is very much higher than that it was intended as /s/, and this would predict that a probability-matching listener perceives such spectral-mean values as /ʃ/ 100 percent of the time. Nevertheless, Figure 8 shows that in this region the learner's perception grammar varies between perceiving /ʃ/ and /s/. This imperfection arises because the learner has not heard enough peripheral tokens to drag the two curves apart in this region.⁸ The same holds for the right periphery of the auditory space (around 9500 Hz).

⁷ In this specific case with only two categories, the probability could be computed directly (rather than just estimated) by the formula in Boersma (1998: 331). However, this would not work for the more complicated case of Figure 17, where each spectral-mean value involves three constraints.

⁸ One might think that considerations of auditory distance (1500 Hz is closer to the centre of the /ʃ/ category than to the centre of the /s/ category) predict that real listeners always classify [1500 Hz] as /ʃ/. However, the uncertainty that our virtual listener displays around 1500 Hz may well be

It is interesting to see for which spectral-mean values the probabilities are maximal. The probability-matching criterion would predict that the probabilities would be maximal in the regions near 1000 and 10000 Hz, but the scarcity of such tokens has moved the point of maximal separation quite far toward the centres of the two categories, i.e. towards 4000 and 7000 Hz. The spectral-mean values for which the two curves are farthest apart are approximately 3900 and 7100 Hz. These locations can thus be said to represent the least confusable spectral-mean tokens and will turn out to be relevant when we consider how the listener of Figure 7 will pronounce /ʃ/ and /s/ herself.

5.5 The near-optimal English listener's preferred production: the prototype effect

In a *bidirectional* model of phonetics, the cue constraints and rankings that the listener uses in perception are also used by her in phonetic implementation. When implementing an articulation, the cue constraint *[1000 Hz]/s/ is read as “an /s/ should not be produced with a spectral mean of [1000 Hz]”, and so on.

In (4) we see how the virtual learner of §5.4 would now produce an /s/, if only the cue constraints and rankings of Figure 7 were involved.

(4) *A production tableau with cue constraints only*

/s/	*[7200] /s/	*[7000] /s/	*[7100] /s/
[7000 Hz]		*!	
☞ [7100 Hz]			*
[7200 Hz]	*!		

The candidate [7100 Hz] in (4) wins because the curve of the cue constraints for /s/ in Figure 7 is lowest at this value (which is also where the two curves are maximally separated). If we compare this phonetic output to the token with the highest frequency in the distribution of Figure 6, which has 7000 Hz, we can see that our speaker produces an /s/ that has shifted slightly from the /s/ that is most commonly produced by her surrounding.

We now show that the shift is actually larger than the 100 Hz seen in (4). As a result of the evaluation noise of Stochastic OT, the winner in (4) will not always be 7100 Hz. Instead, the winner will be the spectral-mean value whose cue constraint (for /s/) will happen to be lowest-ranked when evaluation noise is added to the ranking of each constraint. If we apply the same evaluation noise as in perception, namely with a standard deviation of 2.0, all spectral-mean values with a ranking not much higher than that of *[7100 Hz]/s/ are also likely to win. In Figure 7 we see that the curve of *[x Hz]/s/ has a low plateau in the whole region between, say, 6500 Hz and 8200 Hz, and these are all likely to win. The right-hand side of Figure 9 shows the distribution of spectral values that we obtain by running the category /s/ 1 million times through production tableaux with the rankings of Figure 7.

realistic: from Escudero & Boersma (2004: fig.3) it can be seen that Southern British English listeners have trouble classifying [ɛ] as an instance of the auditorily closer category /ɪ/ rather than of the remoter category /i/.

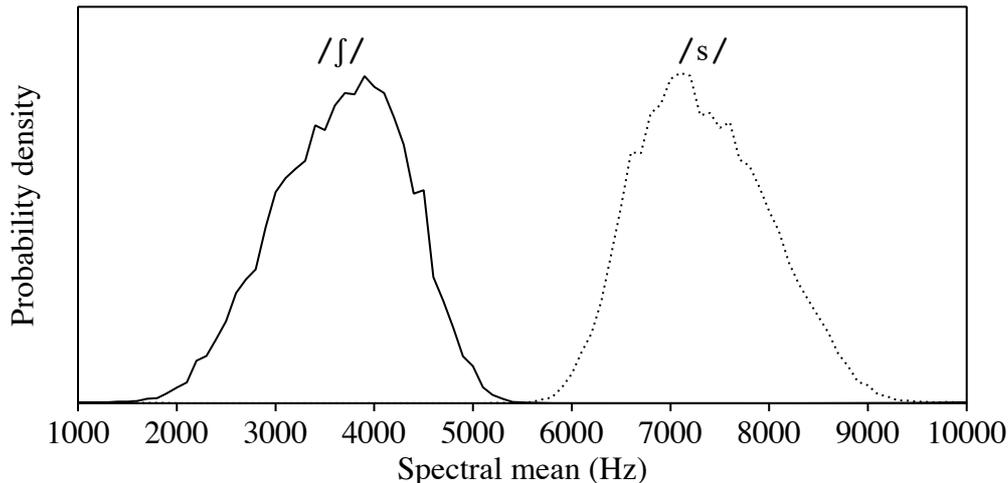


Figure 9

Output distributions of the inverted perception grammar (distributions of ‘prototypes’).

The average spectral mean of the /ʃ/ category is 3664 Hz, that of the /s/ category 7329 Hz. The size of the shift towards the periphery is therefore approximately 330 Hz.

What is behind this shift? From the discussion at the end of §5.4 we glean that a probability-matching learning algorithm tends to lower farthest the constraints for values that are least likely to have been intended by the speaker as anything else than the dominant category in that region. Very roughly, the learning algorithm causes cue constraints to end up ranked lowest in auditory regions where the learner has heard the largest number of least confusable tokens. Teleologically speaking, our speaker prefers to produce a /s/ that is more peripheral than the average token that she has heard herself, because the auditory distance of such a token from the competing /ʃ/ is larger. What we observe here is the so-called *prototype effect* in OT (Boersma 2006). This phenomenon has been attested in experiments in the laboratory: when speakers of a language are asked to choose the best auditory instance of a sound category, they choose a more peripheral token than they would actually produce themselves (Johnson, Flemming & Wright 1993), apparently because they choose the token that is least likely to present any other category than the one requested. It is important to note that although this explanation is couched in teleological terms (as if the listener-speaker knows about this), the underlying mechanism is not explicitly goal-oriented at all (and does not even know about auditory distance): the prototype effect occurs in OT simply because people employ in production the same cue constraint rankings that have optimized their prelexical perception.

The conclusion must be that learners will end up preferring more peripheral tokens than they have heard on average in their language environment. This would predict a sound shift if real learners really did that. But the phonetic implementation process of real speakers is not determined solely by their cue constraints. Articulatory considerations will be seen to keep the prototype effect within bounds.

5.6 Really speaking involves more constraints: the articulatory effect

Something is wrong with the assumption in §5.5 that phonetic implementation involves cue constraints alone. In real production the speaker is restricted by articulatory constraints as well. For instance, we can define the constraint *[7200] as the articulatory constraint whose ranking reflects the articulatory effort associated with producing a spectral mean of 7200 Hz.

Since more peripheral auditory values tend to be harder to produce (cf. §2.1, §4), their constraints must be ranked higher than those for more central auditory values.

In the production of a sound, articulatory constraints interact with the cue constraints. This is illustrated with the production grammar in Figure 10, where articulatory constraints are drawn in thick grey. The cue constraints have the same rankings as in the perception grammar in Figure 7, and are again represented as solid and dotted curves.

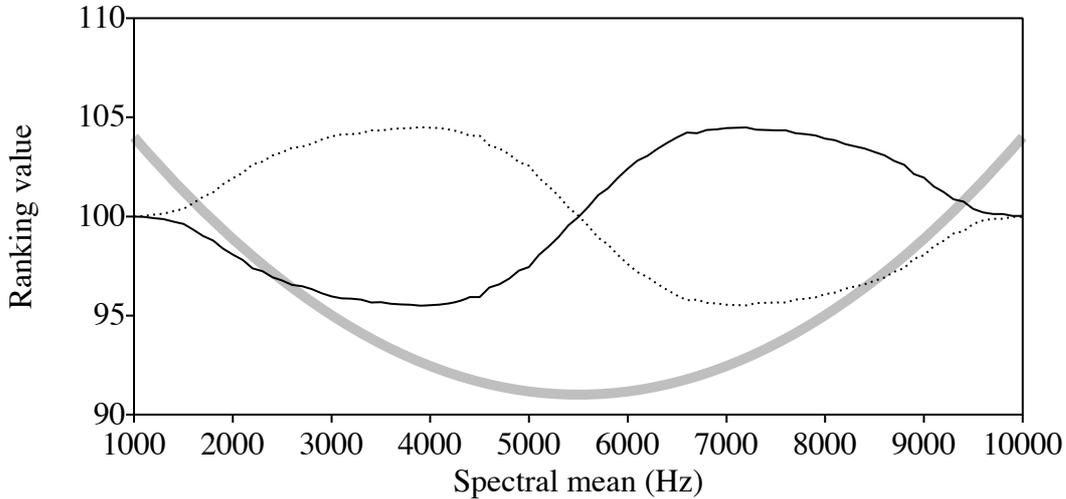


Figure 10

A production grammar for the two sibilant categories in English.

Following Kirchner (1998), we simplifyingly assume that the ranking of articulatory constraints is fixed and not influenced by language-specific learning.⁹

The interactions of articulatory and cue constraints in production become clear with the example of tableau (5), where our learner tries to articulate /s/.

(5) A production tableau with cue constraints and articulatory constraints

/s/	*[7200] /s/	*[7200]	*[7100]	*[7000] /s/	*[7100] /s/	*[7000]
☞ [7000 Hz]				*		*
[7100 Hz]			*!		*	
[7200 Hz]	*!	*				

The candidate [7100 Hz], which won in the tableau without articulatory constraints in (4), no longer wins, because its articulation involves more effort than that of [7000 Hz], which is the new winner (note that in this tableau *[7100] outranks *[7000]/s/ despite the fact that in the region of 7000 and 7100 Hz the articulatory curve in Figure 8 lies below the cue curve for

⁹ A more realistic model would involve articulatory learning, which should lower the ranking of articulatory constraints for spectral-mean values that the speaker has been practicing (Boersma 1998: ch.14).

/s/; there is no contradiction: in stochastic OT, constraints that are as closely ranked as these two will in a non-negligible fraction of the evaluations be ranked in the opposite order).

While the bidirectional use of cue constraints cause the categories to drift apart auditorily (§5.5), the presence of the articulatory constraints checks this expansion and drives the production distributions back towards the centre of the spectral-mean continuum. Figure 11 shows the production distributions, estimated by running each sibilant category 1 million times through the grammar of Figure 10 with an evaluation noise of 2.0.

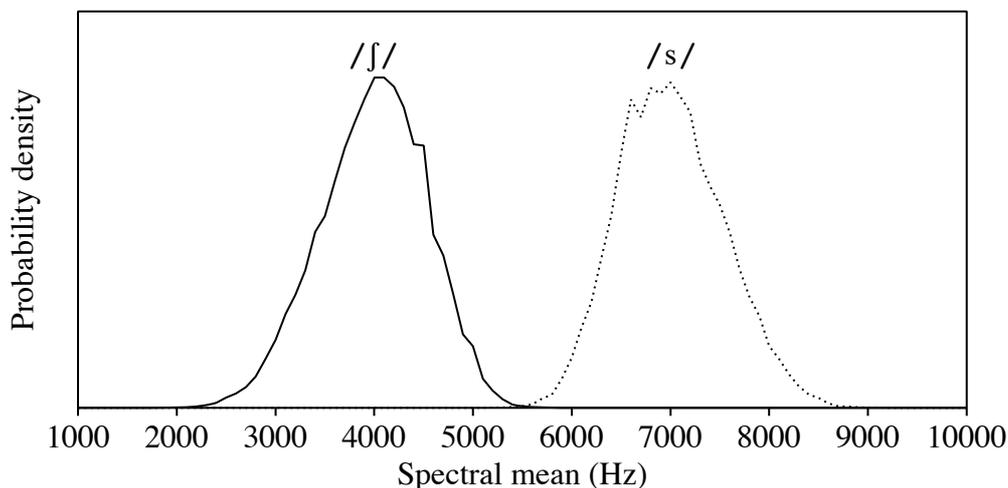


Figure 11

The production distributions for the two sibilant categories in English.

The average spectral mean of the /ʃ/ category is 3991 Hz, that of the /s/ category is 7007 Hz. When we compare this with the averages of Figure 6 (4000 and 7000 Hz) and Figure 9 (3664 and 7329 Hz), we conclude that the articulatory constraints have effectively prevented the categories from drifting apart auditorily: our English learner produces the same average spectral means as her parents. Apparently, the *articulatory effect* has cancelled out the prototype effect. This balance of powers was first noted and modelled by Boersma (2006).

There is an important difference between the distributions in Figure 6 and Figure 11. In Figure 6 the standard deviations of the two distributions were 700 Hz, and in Figure 9 they were 672 and 676 Hz. In Figure 11, however, the standard deviations are only 531 and 528 Hz: the distributions are narrower. Apparently, the articulatory effect caused, beside the shift, an *entrenchment* of the produced distribution when compared with the distribution heard. This sharpening is an important property of our model, which will be seen to help to make the distributions stable over the generations (§5.7) and to compare the model favourably to other models (§7.4).

The near-exact cancelling of the articulatory and prototype effects is a direct result of our choice of parameters: knowing that English has a stable sibilant system, we chose the height of the articulatory curve to be the one in Figure 10, not a higher or lower one. In this sense, the parameters of the model have been established on the basis of language data. Therefore we have to question the equal sizes of the prototype effect and the articulatory effect, and will do so in §5.8 and §5.9. First, however, we show that the language simulated in §5.4 through §5.6 is indeed stable over the generations.

5.7 Simulating sound change: a stable language

Are the average spectral means of the English sibilants, namely 4000 and 7000 Hz, indeed stable over the generations, or do they slowly drift? We test this by simulating the acquisition of the two sibilants over nine more generations.

Some care has to be taken in how the produced spectral-mean values of the first generations of learners are fed to the second generation. We cannot transmit the distribution of Figure 11 directly from speaker to learner, because the only variation in the spectral-mean values of Figure 11 is due to *decision noise* in production. That is, the standard deviation of 530 Hz only reflects the evaluation noise in the production tableaux. The listener will be confronted with this source of variation (and not be able to normalize it away), but also with some non-normalizable between-speaker variation (because real learners will hear multiple speakers), some random variation within the speaker's muscle system (which is independent from the speaker's evaluation noise), the background noise in the air, and the noise in the learner's ear. We represent these influences on the spectral mean together as a *transmission noise* with a standard deviation of 500 Hz. More precisely, the next generations of learners will be presented with a token distribution that is a convolution of the production distribution of Figure 11 and a Gaussian curve with a standard deviation of 500 Hz. The result of this convolution is in Figure 12, which represents the input to generation 2. The standard deviations are approximately 730 Hz.

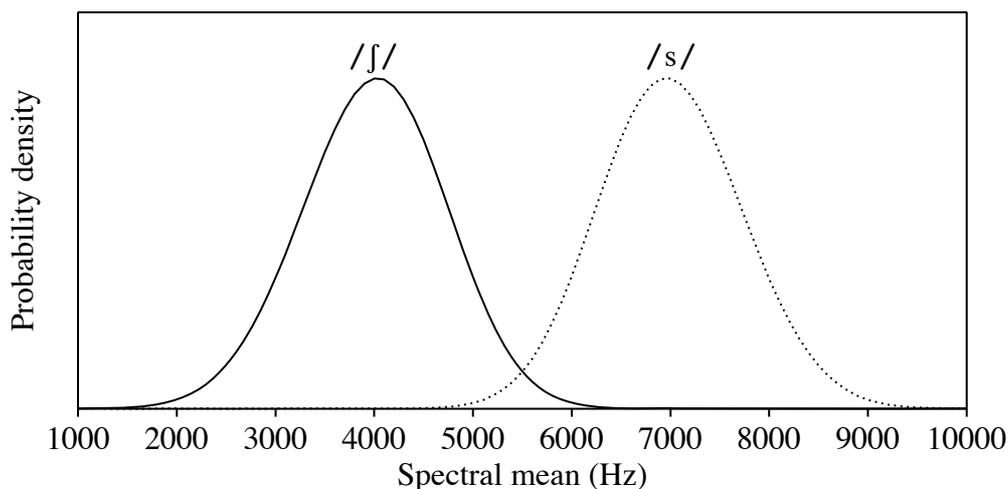


Figure 12

The sibilant environment of the second generation of learners.

Our second generation is thus represented by a learner who acquires a perception grammar in the way described in §5.4, on the basis of the input distribution given in Figure 12 (we again assume that the learner can distinguish the categories from each other). This learner then turns into a speaker by including articulatory constraints, as described in §5.6. We then feed tokens drawn from the output distribution of this speaker, convolved with 500-Hz transmission noise, to a new learner, our third generation. This learner acquires again a perception grammar and a production grammar; her convolved output distribution is used as input to the fourth learner, and so on.

In total we simulate this acquisition process for 10 learners in a row, with the simplifying assumption that every learner stands for a whole generation of speaker-listeners. Every learner

receives 1 million /ʃ/ and 1 million /s/ tokens, and has a plasticity of 0.01 and an evaluation noise of 2.0.

The English spectral mean values of 4000 and 7000 Hz turn out to stay constant over the generations, and the standard deviations stabilize at about 735 Hz. This is shown in Figure 13, where the black curves connect the single-speaker averages and the grey areas represent the standard deviations of distributions like those in Figures 6 and 12 (i.e. including transmission noise).

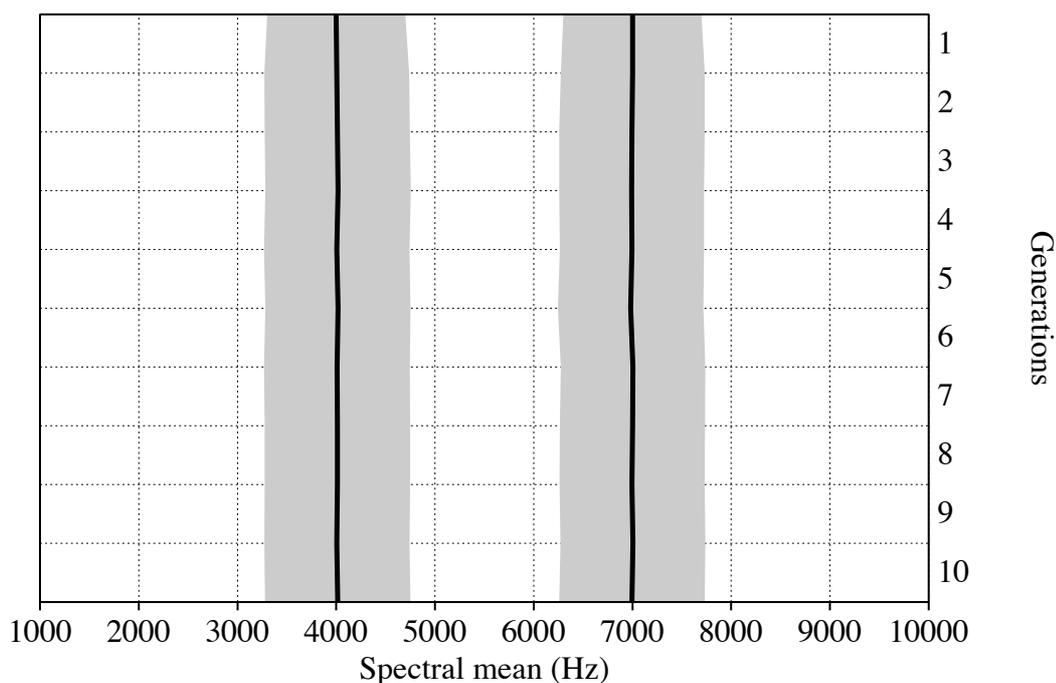


Figure 13

The spectral means of the two English sibilants and their stable learning over 10 generations.

Our simulated English (Figure 6) turns out to be boringly constant (Figure 13). There is no change between the output of generation 1 and the output of generation 10. The next question is whether a language with different auditory distributions than Figure 6 is equally stable, or whether it instead drifts away from its initial distributions.

5.8 Simulating sound change: a language with an exaggerated contrast is unstable

Here we simulate the learning of a language with a much more extreme two-sibilant contrast than English, i.e. [ʃ] versus [ʂ], with spectral means of 2500 and 8500 Hz. Except for these initial distributions, everything else (including the shape and height of the articulatory constraint curve of Figure 10) is the same in this simulation as in the one of §5.4 through §5.7.

Figure 14 shows the result. For this ‘exaggerated English’, the second generation more or less learns the oversized range, although they reduce it to a much more moderate contrast of 3500~7500 Hz. Apparently, the articulatory effect outweighs the prototype effect for this generation; this comes to no surprise, regarding the height of the articulatory curve in Figure 10 in the regions of 2500 and 8500 Hz. The third generation has already shifted the system towards an unmarked articulatorily-perceptually balanced [ʃ] and [s]. Within two generations, the learners have changed the exaggerated English into plain English. The conclusion is that

the articulatory and prototype effects can act independently, and work together to establish, non-teleologically, an optimal balance between articulatory ease and auditory contrast.

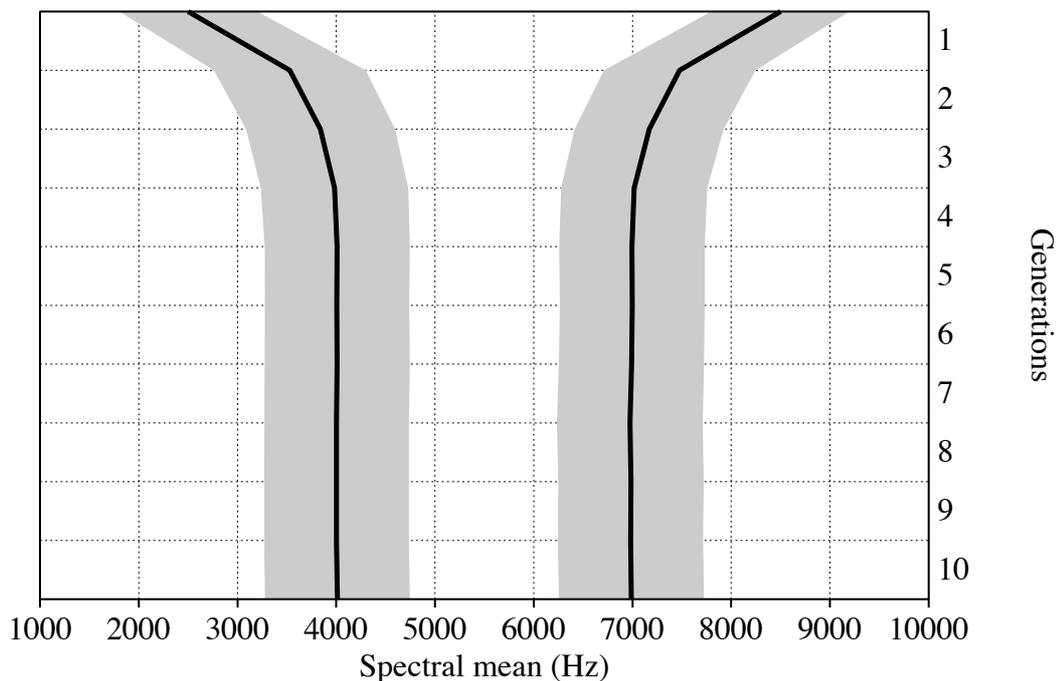


Figure 14

The development of the two sibilants in ‘exaggerated English’ over 10 generations.

5.9 Simulating sound change: a language with a confusable contrast is unstable

If the optimal balance achieved in Figures 13 and 14 is characteristic of two-sibilant inventories in general, then the 4000~7000 Hz inventory should emerge for every possible initial inventory. The two cases that yet have to be investigated in this respect are the case of an initial ‘confusable’ English, which has its categories spaced by only e.g. 2000 Hz, and the case of an initial ‘skewed’ English, where the categories are not positioned symmetrically around 5500 Hz (as they are in Figures 13 and 14).

The two cases are combined in the simulation of Figure 15, a language with initial sibilants at 5500 and 7500 Hz, which is both skewed (one sibilant has a central, the other a high spectral mean) and relatively confusable (the difference is only 2000 Hz rather than the 3000 Hz of the earlier simulations).

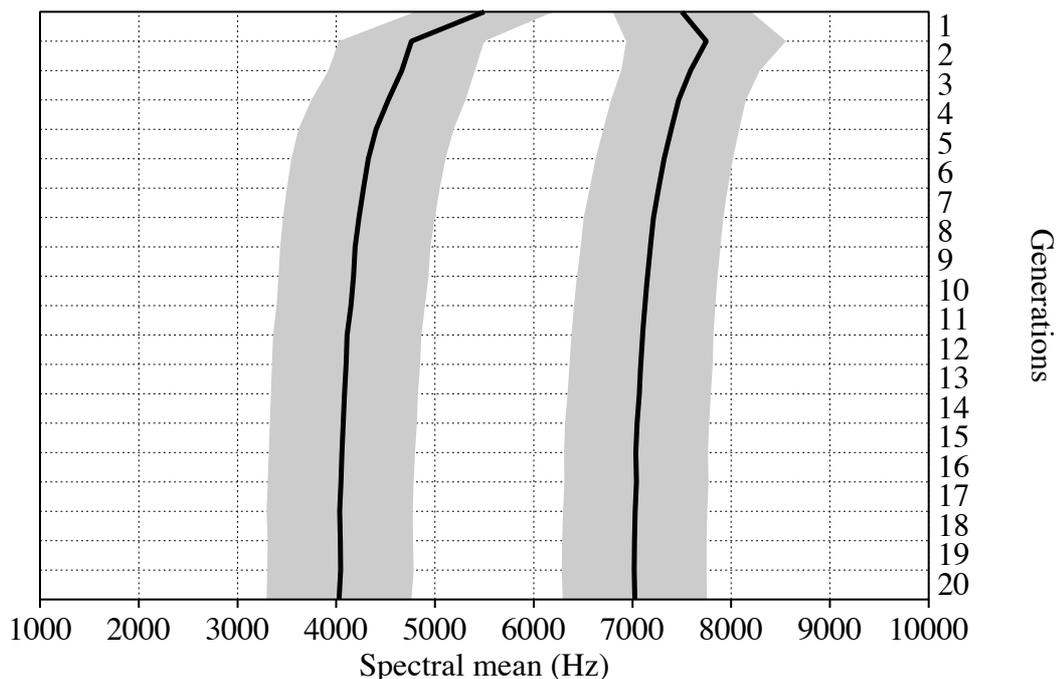


Figure 15

The development of the two sibilants in ‘skewed and confusing English’ over 10 generations.

The first generation of learners moves the spectral-mean values apart (to 4762 and 7748 Hz), immediately reaching an English-like distance of nearly 3000 Hz. Apparently, the prototype effect plays a very fast role here; this confirms the repulsive dynamic nature of the present model and shows that the model will be able to handle phonetic enhancement and chain shifts. The articulatory effect is also seen to play a role, because the amount the lower category moves down (i.e., to the left in the figure) is larger than the amount the top category moves up. This effect works quite slow, however: it takes quite a number of generations to obtain a perfectly symmetric inventory. The cause of the slowness is that the upper category ‘wants’ to move down but that this desire is hampered both by the repulsive force from the lower category and from the fact that the lower category itself experiences an upward bias from the articulatory effect. Nevertheless, a symmetric English-like inventory is eventually reached.

We can conclude from Figures 13 through 15 that our model predicts that independently of the situation in generation one, the inventory always evolves towards the same two auditory values. In other words, all stable languages with two sibilants are like English (or French, or German).

So we see that optimal dispersion indeed happens. Figures 13, 14 and 15 all show the effect of the excluded centre (§2.1): the region around 5500 Hz is avoided in languages with two sibilants. The next step is to look at other kinds of sibilant inventories: which of the dispersion effects will we see?

6. Larger, smaller, and different inventories

In order to find all six dispersion effects discussed in §2.1, the present section discusses inventories with three and four categories (cf. Figure 5). As special cases we also discuss an inventory with a single category and an inventory containing a non-contiguous category.

6.1 The Polish three-sibilant inventory: a chain shift

Polish used to have the sibilants /ʃ sʲ s/ (Carlton 1991). If the dispersion principle is correct, and if the spectral mean is the only auditory cue that distinguishes these three sibilants, such an inventory cannot be stable. The simulation in Figure 16 shows what happens if we start out with a language whose sibilants have spectral means of 4000, 6700 and 7000 Hz (in order to make our point, we have taken these last two closer together than they probably actually were in medieval Polish).

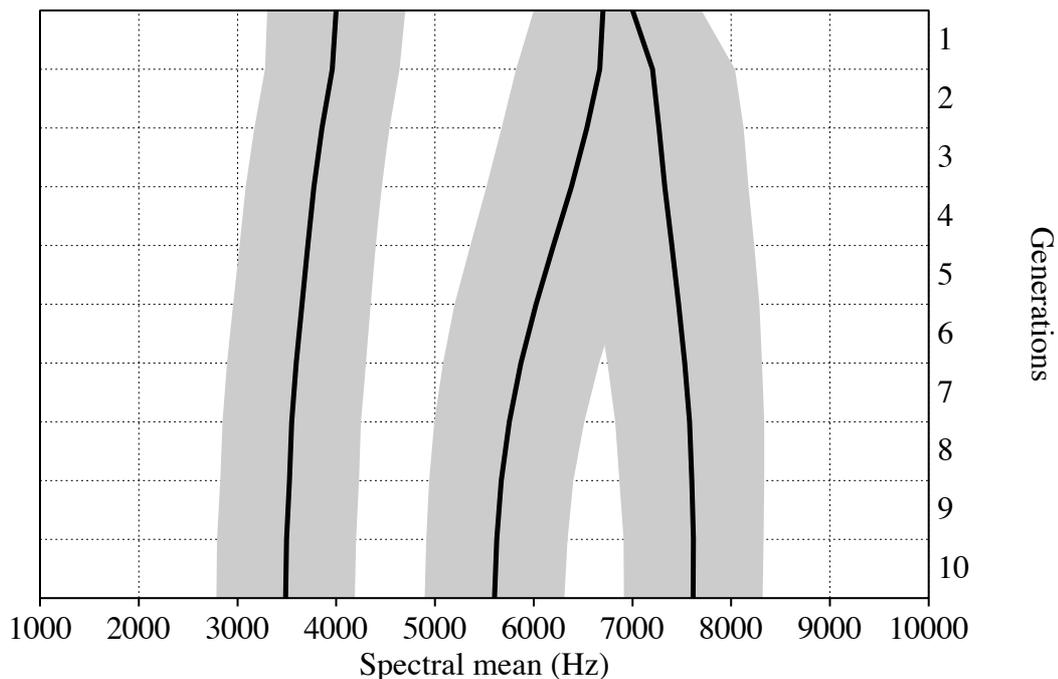


Figure 16

The development of the three sibilants in Polish over 10 generations.

The first striking phenomenon that Figure 16 shows is that the /sʲ/~s/ contrast is phonetically enhanced: the palatalized alveolar (the middle category) lowers its spectral mean beyond 6000 Hz, i.e. into the [ç] region. This is reported to have happened in real Polish in the 13th century (Stieber 1952, Carlton 1991). The second thing that happens is that the postalveolar (the sound on the left in Figure 16) shifts down towards [ʃ], which is reported to have happened in real Polish in the 16th century (Rospond 1971, though this sound is usually transcribed as postalveolar /ʃ/; see the discussion in Hamann 2004). Our simulation confirms Jones' (2001) proposal that this second shift was caused by the first, i.e. that we are observing a contrast-enhancing push chain here: the /ʃ/ category is shifted down as a result of the approach by the lowering /ç/. Furthermore, we can observe the equally contrast-enhancing shift (again proposed by Jones 2001) of the alveolar /s/ towards a more peripheral [ʂ], its present-day location (Puppel *et al* 1977). This simulation thus explains the present Polish sibilant system /ʃ ç ʂ/, which has spectral means not far removed from the 3500, 5500, and 7500 Hz found here (Zygis & Hamann 2003). With these three spectral means in her auditory environment, a Polish learner will acquire the production grammar in Figure 17. We have thus established an account of the Polish sibilant inventory without using dispersion constraints (cf. Padgett & Zygis 2003, who did use dispersion constraints for this example).

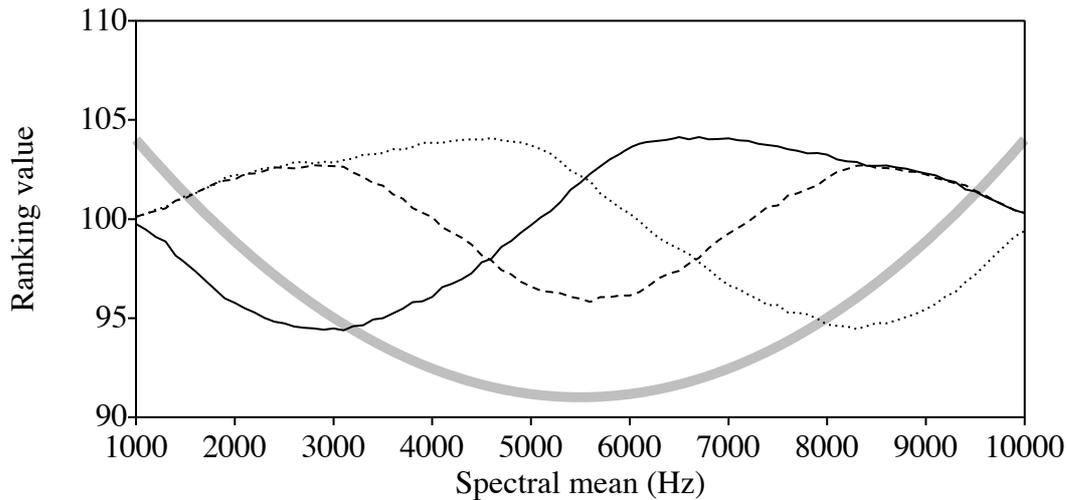


Figure 17

A production grammar for the three sibilant categories in Polish.

When we compare Figure 16 to Figure 15, we see all the dispersion effects discussed in §2.1. First, the centre category (around 5500 Hz) is back in the picture. Second, the dispersion in Figure 16 is even: both pairs of adjacent categories are removed from each other by approximately 2000 Hz. Third, we see that the size of the auditory space is greater when there are three categories than when there are two. Fourth, from the widths of the grey areas we see that the stable variation within a category is somewhat smaller when there are three categories (namely 705 Hz) than when there are two (735 Hz, according to §5.7). Fifth, we have observed a push chain when a lowering mid category caused the bottom category to lower as well. Sixth, we will observe a drag chain if we remove the bottom category from the final state in Figure 16: from the simulation in §5.9 we know that the original mid category will move down to 4000 Hz, allowing the top category to move down from 7600 to 7000 Hz. We conclude that we have faithfully modelled all aspects of auditory dispersion.

6.2 A four-sibilant inventory

The largest inventory we consider is one with four sibilants, like Toda in Figure 5. Independently of where the centres of the categories lie in the environment of the learners of the first generation, the centres of the categories evolve towards values around 3400, 4600, 6400, and 7600 Hz, i.e., the distances between the categories are again smaller than in the three-sibilant case, while the total space taken up has again increased (although very little). The spacing between the categories is somewhat greater in the middle than at the edges, probably because at the edges the limiting effect of the articulatory constraints is greater (this can be seen as a more precise formulation of §2.1, point 3). The final standard deviations are 720 Hz for the two outer categories and 810 Hz for the two inner categories, i.e. a bit larger than in the three-sibilant case.

We conclude that all dispersion effects identified in §2.1 remain valid in larger inventories (except the unexpectedly high within-category variation in the four-sibilant inventory, for which we have no explanation).

6.3 A one-sibilant inventory

The smallest inventory we consider is one with a single sibilant. The first generation of learners already turns up with a category centred at the centre of the auditory continuum, i.e. 5500 Hz, and the standard deviation will be and stay 1700 Hz. The cause of this situation is that the learners do not learn: if the only category is the unspecified sibilant /S/, they will perceive any auditory value as /S/, and therefore never make a mistake. As a result, all cue constraints stay ranked at 100.0, and in production the decision about which auditory value to pronounce is determined partly by which cue constraint happens to be lowest ranked, partly by the articulatory constraints.

The reader may object that the predicted one-shot shift to the centre is unrealistic, yet this is what we predict will happen in the absence of other phonological entities. In practice it will be very difficult to find such a situation: the single (retracted apico-alveolar) sibilant of Spanish, for instance, must cope with the existence of a rather strident /θ/ in the same language.

6.4 Can non-contiguous categories be learned?

The present model does not involve any representation of auditory distance. That is, the learner represents nothing more than 91 values along the spectral-mean continuum, and does not necessarily know that they are ordered in a natural way. For instance, the virtual learner never needs to know that the spectral mean associated with the 17th value along the continuum (i.e. 2600 Hz) lies in between those of the 16th value (2500 Hz) and the 18th value (2700 Hz). As a result, the connections of our virtual learner can easily represent an inventory where category 1 has spectral means around 3000, 5000, and 7000 Hz, whereas category 2 has spectral means around 4000, 6000, and 8000 Hz (as was mentioned in §5.2). The question naturally arises, however, whether such discontinuous categories are not only representable, but learnable as well.

The answer turns out to be that discontinuous categories are *semi-learnable*: if the learner's environment contains a discontinuous category, then the learner's final category may still be discontinuous, but less so than the one in her environment, and within a number of generations the language will have changed into one with exclusively contiguous categories.

An example is shown in Figure 18. The initial inventory has a simple (monomodal) category (which we arbitrarily label /ʃ/) centred at 5500 Hz with the usual relative frequency of occurrence of 50%, plus a bimodal category (which we label /s/) with a peak centred at 7500 Hz with a relative frequency of 37.5% and a peak centred at 3500 Hz with a relative frequency of 12.5%; all three peaks have a standard deviation of 700 Hz. Thus, both categories are equally likely, and the upper part of /s/ is three times more likely than the lower part.

Figure 18 shows that the following generations continue to associate the /s/ category with two peaks. However, the lower peak shrinks slowly, and has almost disappeared after generation 11. After 20 generations, the English inventory has arisen again.

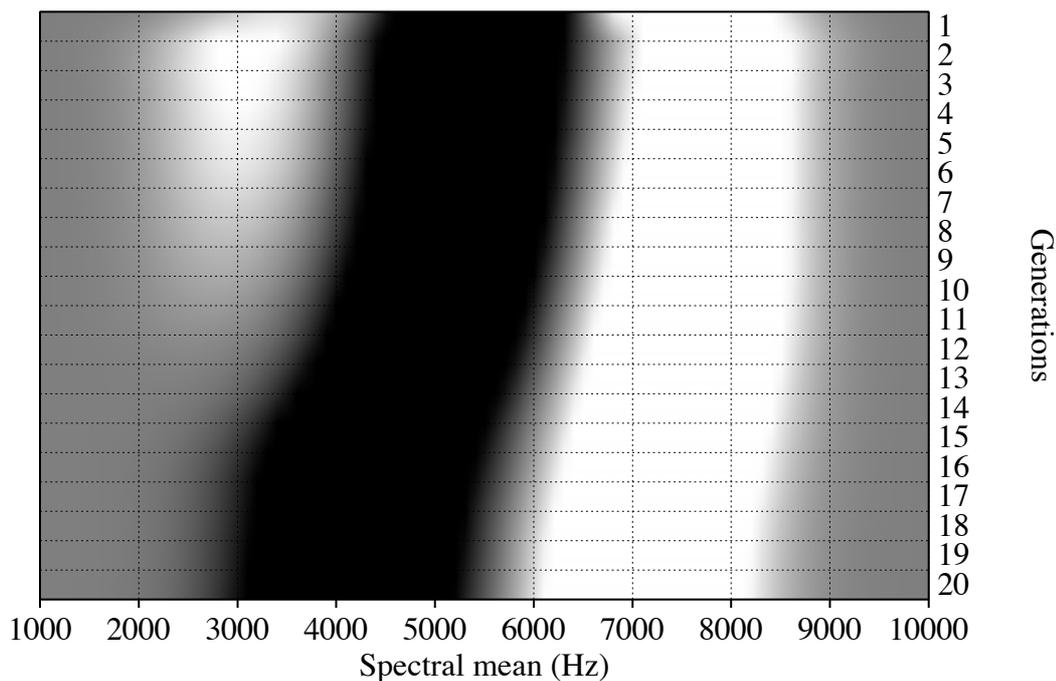


Figure 18

The demise of a bimodally distributed category. Black = /ʃ/, white = /s/.

We conclude that in our model non-contiguous categories are learnable, but not stable over the generations. The ultimate cause of the development into monomodal distributions can be explained with Figure 18 in the following way. The fact that the upper peak of /s/ is taller than the lower peak causes an asymmetry in the regions of confusion. That is, the region of confusion between the lower peak and /ʃ/ lies somewhat further from the centre of the continuum (namely around 4200 Hz, which is 1300 Hz away) than the region of confusion between the lower peak and /s/ does (around 6600 Hz, which is only 1100 Hz away). As a result, /ʃ/ will shift down (i.e. to the left), which then pushes the lower peak of /s/ down into an articulatory more effortful region (push chain) and allows the upper peak of /s/ to come down into a less effortful region (drag chain). As a result, the articulatory bias against the lower peak of /s/ becomes greater than the articulatory bias against the upper peak, and this causes a slight preference in production for selecting tokens from the upper peak. This process is self-reinforcing, because it moves the boundaries between the three categories down.

We do not know yet how to assess our predicted relative learnability of bimodal distributions. If there exist mechanisms that cause such distributions to arise, we must be able to observe them in real languages because such distributions are predicted to take ten generations to become monomodal. Whether actual languages do have this kind of transitory allophony in sibilants remains to be seen.¹⁰

¹⁰ If we are allowed to speculate: perhaps Dutch is becoming a case in hand. The introduction of the relatively new and somewhat marginal alveopalatal sibilant /ç/ (usually transcribed as /ʃ/) into an existing inventory consisting solely of an (auditorily equally central) flat laminal alveolar sibilant (usually transcribed as /s/) may cause a split in the population between varieties of /s/ with higher and those with lower spectral means than /ç/. Unfortunately, the present paper has no room to test this speculation.

7. Discussion

Our simulations within a bidirectional phonetics model realistically predict that a language with one, two, three or four sibilants automatically evolves towards a *dispersed* system, i.e. one that has single-peaked categories equally spaced along the auditory spectral-mean continuum. The end result of such an evolution is independent of the spectral means of the categories in the first generation: given the number of categories, the resulting final categories will always be monomodal and have the same averages and standard deviations. Another thing our simulations have modelled realistically is the diachronic development of the sibilant inventory of Polish.

Our approach reconciles the standpoint of innocent misapprehension with that of auditory dispersion: speakers are not goal-oriented, but at the same time sound change tends to minimize perceptual confusion. Sound change at the level of the language learner is thus non-teleological, whereas at the abstract level of the observed language it is teleological. In this section we discuss the required assumptions and parameters of our model and compare them with those of earlier models.

7.1 Assumption 1: multiple levels of representation

The reduction to non-teleological underlying mechanisms is made possible by going beyond OT's usual two-level unidirectional grammar model, in which phonetic and phonological representations would have to be mixed (as in Flemming's 1995 OT version of Dispersion Theory). To unmix the phonetic and phonological representations, they have to be regarded as separate, and as equally important. This is accomplished by the model in Figures 2 and 3, in which the phonological surface form and the phonetic form are distinct representations connected by OT constraints in a way similar to how faithfulness constraints tend to connect the underlying form and the phonological surface form. The resulting total model is a multi-level OT with at least three levels of representation: underlying form, surface form, phonetic form. And if we want to distinguish more rigorously between auditory form and articulatory form than we have done here, and to distinguish between form and meaning in the lexicon (Apoussidou 2006), we arrive at the five-level model in Figure 19.

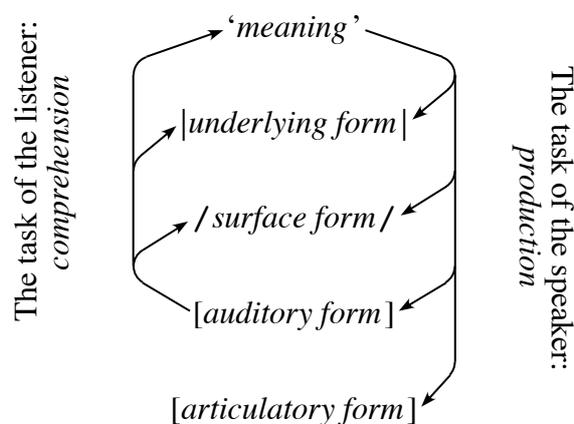


Figure 19

The bigger picture.

The arrows in Figure 19 have been drawn in parallel to show that although the connections between the representations are local (e.g. cue constraints connect auditory to

surface form and faithfulness constraints connect surface to underlying form, but there are no direct connections from auditory to underlying form), the optimizations are performed globally. In our case of sibilant production, for instance, cue constraints (between auditory and surface form) interact with articulatory constraints (at articulatory form); in another case (Boersma 2005), cue constraints have been shown to interact with faithfulness constraints as well.

The full model thus rigorously separates the two discrete phonological representations from the two continuous phonetic representations, and is therefore compatible with many areas of current phonological investigation. At the same time, this model allows the phonology and phonetics to interact through global evaluation, thus reconciling the observation that phonological elements seem to be discrete with the observation that continuous-phonetic considerations seem to influence phonological behaviour.

7.2 Assumption 2: bidirectionality

Beside featuring multiple levels, the model has to be bidirectional, in the sense that the same constraints are used for modelling the speaker and the listener: the prototype effect arises because the constraint ranking that optimizes the mapping from phonetic to surface form is reused in production (Boersma 2006).

Within Optimality Theory, bidirectionality has been proposed before by Blutner (2001), Zeevat & Jäger (2002), and Jäger (2003). These proposals (for cases of competition in semantics and pragmatics) try to optimize the listener and the speaker at the same time. The paper by Jäger (2003) is very close to the present paper in its general methodology of simulating acquisition and evolution and in its line of thought; for instance, Jäger presents several cases of language types that are both representable in UG and learnable, but unstable over the generations. However, Jäger's Bidirectional Gradual Learning Algorithm relies on a slightly teleological feature of evaluation in production: every candidate form in a production tableau has to be *hearer-optimal*, i.e. if taken as the input to a comprehension tableau (with the same rankings) it should be mapped to a meaning identical to the input of the production tableau.¹¹ This explicitly listener-oriented evaluation procedure thus militates against ambiguous (i.e. poorly 'dispersed') forms in production, and Jäger relies on it for establishing the diachronic emergence of pragmatic case marking (which enhances the semantic contrast between subject and object). It would be interesting to investigate whether our arguably simpler procedure (optimize comprehension only, then just speak) would be able to handle the complex cases that Jäger discusses.

Outside OT, bidirectionality has been proposed for an exemplar model of phonology. In exemplar theory (Johnson 1997, Lacerda 1997), the listener stores in her memory the actual phonetically detailed tokens ('exemplars') she hears, together with their category labels. Pierrehumbert (2001) proposes that the listener will subsequently use the same exemplars in production. It has been claimed by both Lacerda (1997: 27) and Pierrehumbert (2001: 143) that a prototype effect similar to the one we derived in §5.5 will arise in listener's goodness

¹¹ A similar criticism can be raised against Boersma's (1998) production model, which relies on a *control loop*, in which every candidate in production is a triplet of articulatory, auditory and surface forms where the auditory form is optimal given the articulatory form and the surface form is optimal given the auditory form.

ratings (although neither Lacerda nor Pierrehumbert provide the simulations to prove this¹²). Blevins (2004: 285–289) proposes that the shifted ‘prototypes’ (i.e. the ‘best’ exemplars in goodness ratings) will be used in production as well (p. 288), ultimately leading to chain shifts. That this is indeed possible in exemplar theory was proven by the simulations by Wedel (2004: 140–169, 2006: 261–269), which use a device for detecting ambiguous tokens, i.e. tokens whose auditory values lie in the regions of overlap between two categories. By refusing to store the ambiguous tokens as exemplars at all, or by allowing them to be stored in the incorrectly perceived category (rather than in the category intended by the speaker, as happens e.g. in our lexical supervision), Wedel derives the required repulsion of categories. The similarity between Wedel’s simulations and ours is that this prototype effect is caused in perception by cross-category competition in regions of overlap. When we compare the assumptions required for this competition effect in Wedel’s and our simulations, we see that in our case the competition has been at the core of the architecture of the theoretical framework from its inception, namely in the combination of the general evaluator (EVAL) that handles OT production (Prince & Smolensky 1993) as well as OT perception (Boersma 1997) with the general error-driven mechanisms that handle learning in OT production (Tesar 1995) as well as in OT perception (Boersma 1997). In Wedel’s case, by contrast, the competition is achieved by a separate ambiguity-detecting device with the specific purpose of maintaining contrast. Whether this device will ultimately turn out to be as generally useful in exemplar theory as EVAL and error-driven learning are in Optimality Theory, is an interesting possibility that we cannot evaluate yet.

In all, our model seems to be the one in which the prototype effect in perception and its repulsive effect in production come about most naturally. Given bidirectional OT with optimization of comprehension, we would in fact need complex additional machinery if we did *not* want categories to repel each other. The fact that this simplest possible bidirectional OT model already exhibits this effect, an effect that arguably contributes to the success of the species that have it, suggests the idea that biological evolution has selected this model as a general decision-making mechanism.

7.3 Assumptions required to go beyond the limitations of the present paper

Our present simple implementation of the model comes with several limitations.

Consider the problem mentioned in §5.4 that it is unrealistic to assume an initial state with fully random prelexical perception together with perfect lexical representations. In a more realistic simulation we will have to take into account the likely fact that the contrast between the categories /s/ and /ʃ/, which has to have been acquired before these categories can be used in lexical entries, has emerged in the learner as the result of prior distributional learning, which should have given a non-trivial initial ranking of the cue constraints. An explicit proposal was provided by Boersma, Escudero & Hayes (2003). We will also have to take into

¹² Lacerda did do some computations, but only for the perceptual-magnet effect. Pierrehumbert did run simulations for production, but did not implement the reciprocal inhibition that would have been necessary to illustrate the prototype effect in perception, let alone in production; in fact, Wedel (2004: 195) points out that in Pierrehumbert’s simulations cross-category blending in assembling production targets will eventually cause all categories to merge into one. Pierrehumbert’s simulations are therefore an instance of the clustering algorithms mentioned in §2.1, point 6. Following an argument by Wedel (2006: 259–261), one can see that the same is true of Lacerda’s (1997) computations.

account the likely fact that lexical representations are acquired in concurrence with prelexical perception. An explicit proposal was provided by Apoussidou (2006). Both refinements will require a comprehensive simulation that is far outside the scope of the present paper.

Another limitation is that our model makes no explicit provisions for diachronic merger. This is because we assumed that even if two categories were very close to each other, as in the initial situation of Figure 16, the learner was given correct labels by her lexicon. However, a beginning learner does not know beforehand how many sibilants her future language has, and therefore at least has to rely on a stage of distributional learning, in which she establishes the number of categories: if the distributions of two adjacent adult categories overlap too much, these distributions may not form separate peaks and the infant may posit a single category instead of two (e.g. Boersma, Escudero & Hayes 2003); if she does, a merger has occurred. The comprehensive simulation mentioned in the previous paragraph will take care of this situation. Thus, the present model, combined with the independently needed earlier stage of category creation, accounts for both chain shift and merger.

Finally, we limited our discussions to fixed locations of category centres, rather than acknowledging the fact that speakers adapt their auditory spaces to pragmatic circumstances on the fly. It does not seem very difficult to include such facts, though. Under circumstances of extra articulatory effort (e.g. fast speech), for instance, speakers could indiscriminately raise the rankings of all articulatory constraints by the same amount (Boersma 1998: 275). The auditory values will then become less peripheral immediately (i.e. without learning), just as in the real world. Under circumstances of extra clarity (e.g. addressing a crowd), speakers could raise the rankings of all cue constraints, as well as perhaps those of all faithfulness constraints (Boersma 1998: 275), sensorimotor constraints, and lexical constraints. The auditory values are then predicted to become more peripheral immediately (i.e. without learning), just as in the real world. In this way, our model can account for instant adjustments in hyper- and hypospeech (Lindblom 1990). In the same way it can be connected well to the way in which Boersma (1998: 208–215) modelled in OT the dependence of auditory forms on stress and context.

7.4 Assumptions not required

In order to be able to assess the position of the present account within the entire set of accounts of auditory dispersion, we have to compare the assumptions of those other accounts with those of ours. We argue that beside the multiple representations and bidirectionality mentioned in §7.1 and §7.2, the model defended here requires none of the special assumptions of the other accounts.

For instance, *transmission noise* is not represented in the mind or brain of the speaker-listener. It comes for “free” in the transmission between the speaker’s production decision and the listener’s auditory-phonetic input. Likewise, *prototypes* are not represented in the present model, although they could be computed by tableaux such as (4). In the same way, *category boundaries* are not represented either, nor are any *statistics* such as the averages and standard deviations of the spectral-mean distributions. The *distributions* themselves are only very indirectly represented in the rankings of the cue constraints. Finally, the speaker-listener does not need to store *exemplars* of every category, because the frequency effects that exemplar theory ascribes to the multiplicity of exemplars tend to derive automatically from the rankings of the many constraints.

Although we use a single auditory continuum in this paper, the model can be applied easily to multiple dimensions, and scales well (i.e. linearly) with the number of dimensions.

For instance, the simulations can be done on a vowel space that has two auditory dimensions (auditory height or first formant, and auditory backness or second formant). If we divide these two auditory continua into 100 points each, we require only 200 cue constraints per category (Boersma & Escudero to appear [2004]). By contrast, models that store distributions directly scale poorly (i.e. exponentially) with the number of dimensions: for instance, they would require 10000 points to maintain a joint distribution for the two continua. We cannot yet assess how the exemplar models by Lacerda (1997) and Pierrehumbert (2001) scale with increasing dimensionality, because they only address single continua.

For the stability of the *standard deviation* of a category over the generations, nothing has to be assumed: the stability is an automatic consequence of the balance between the entrenchment effect of the articulatory constraints (§5.6) and the transmission noise. By contrast, exemplar theory has to invoke special measures to keep the standard deviation stable, as noted by Pierrehumbert (2001: 149–152): if speakers just randomly choose from among their previously stored exemplars, the transmission noise will soon increase the variation between the exemplars of the category; as a consequence, Pierrehumbert has to propose that speakers use an average of multiple exemplars, and this leads to the required entrenchment. Wedel (2006) calls this averaging *within-category blending inheritance*; the present model, by contrast, can get by with choosing a single existing auditory value.

Our model does not represent or compute *auditory distance*. In the exemplar model by Pierrehumbert (2001), listeners need auditory distances to compute the degree to which an existing exemplar is activated (p. 141), and speakers need it to compute the contribution of various exemplars in establishing the ‘entrenchment’ that has to undo the effects of the transmission noise (p. 149). A very far-reaching representation of auditory distance is needed in OT Dispersion Theory (Flemming 1995) to assess violations of MINDIST. For instance, a language user with categories A, B and C has to compute all distances A-B, B-C and A-C, and subsequently to decide which of these is the smallest. In our model, a possible unwanted side effect of not representing auditory distance, namely the learnability of the non-existent or very rare non-contiguous categories, is counteracted by the inherent instability of such categories over the generations (§6.4).

Finally, and most importantly, the present model does not represent any devices explicitly oriented at the goal of improving dispersion, such as the dispersion constraints MINDIST (Flemming 1995), SPACE (Padgett 2003ab), or \mathcal{D}_n -P (Sanders 2003). Whether this means that dispersion constraints are superfluous for phonological theory depends on the question whether they can be used for other things beside evaluating inventories. For instance, dispersion constraints have been used to describe diachronic phonetic enhancement (Sanders 2003: 123). Since phonetic enhancement is also a feature of our model (§5.9, §6.1), dispersion constraints do not seem to be required for modelling such phenomena. An opposite use of MINDIST has been to describe neutralization effects: because of the way MINDIST has been formulated (loosely “small auditory distances between contrasting categories are not allowed”), one way to satisfy it is to neutralize the contrast completely. One such effect, namely unconditional diachronic merger, is handled in our model by the innocent misapprehension that takes place in the infant’s distributional learning stage (§7.3). The other effect, namely conditional merger, involves the phonologically more interesting cases of assimilation and positional neutralization (e.g. Flemming 1995 [2002: 119–151]). The bidirectional phonology and phonetics model generally accounts for such phenomena by

ranking cue constraints over faithfulness constraints (Boersma 2005: 38–39)¹³. A rigorous analysis in these terms of all effects ascribed to dispersion constraints in the literature falls outside the scope of the present paper.

7.5 Sensitivity to the parameter settings

The present model requires five parameters: the shape and height of the fixed articulatory effort curve (Figures 10 and 17); the amount of transmission noise (500 Hz); the plasticity (0.01); the granularity of the spectral-mean continuum (100 Hz); and the number of training data (1 million per category).

The results are not qualitatively sensitive to the exact values of these parameters. Raising the effort curve or steepening it toward the sides will move the emerging category centres further towards each other, whereas raising the transmission noise will move them further apart. Raising the plasticity by a factor of 10 moves the curves of the cue constraints slightly further apart, although this effect can be compensated by reducing the number of training data by a factor of 10.

The results are sensitive to the granularity of the spectral-mean continuum in an unrealistic way. Reducing the granularity from 100 to 10 Hz causes fewer tokens in each spectral-mean value, hence a reduction of the distance by which the cue constraints will move. A more realistic simulation of the continuum would have to take into account the fact that the properties of the basilar membrane in the inner ear are such that when a nerve fiber at a certain frequency is excited, several nerve fibers within a frequency distance of about one-third of an octave will also be excited. As a result, learning will push Gaussian-like dents into the shape of the cue-constraint curves. This correction was taken into account in the simulations by Boersma (1997: 51).

7.6 Are phonological features innate or emergent?

In this paper we have used /ʃ/ and /s/ as arbitrary labels that had no preference for any specific positions along the spectral-mean continuum. Given this arbitrariness, our model seems slightly more compatible with the viewpoint that phonological features themselves are emergent (Boersma 1998, Blevins 2004, Mielke 2004) than with the viewpoint that phonological features are innate. This is because innate features just seem to be in the way of our learning procedures. For instance, if distributional learning creates three peaks, and therefore three categories, along an auditory continuum, it is easier to give these categories arbitrary labels than to associate them to innate feature values, a procedure that would require an additional mapping device. We realize, though, that the controversial issue of innateness versus emergence is too big for the present paper.

8 Conclusion: the innocent emergence of optimal dispersion

The two assumptions of multiple levels and bidirectionality have explained the origins and stability of auditory dispersion over the generations. It turned out that if the auditory category centres are too wide apart, the learner's articulatory effect will be greater than her

¹³ In Boersma's example, the underlying form $|\text{k}\epsilon\text{l}\#\text{?}\text{a}\text{z}\text{a}\text{?}|$ can be realized as the phonological-phonetic output pair $/\text{k}\epsilon\text{l}\text{a}\text{z}\text{a}\text{?}/[\text{k}\epsilon\text{l}\text{a}\text{z}\text{a}\text{?}]$ with deletion of the underlying glottal stop in a position where it is poorly audible (namely after consonants). This is formalized as the ranking of the cue constraint $*[\text{C}\text{?}\text{V}]/\text{C}\text{?}\text{V}/$ above the faithfulness constraint $\text{MAX}(\text{?})$.

prototypicality effect, forcing her to shift her production partly towards a smaller, more naturally dispersed inventory, and that if the auditory category centres are too close, the learner's articulatory effect will be smaller than her prototypicality effect, forcing her to shift her production towards a larger, again more naturally dispersed inventory. Over the generations, every language innocently evolves towards a stable, typologically natural inventory that strikes an optimal balance between articulatory ease and auditory contrast. Everything else being equal, we expect a stable language with two or three sibilants to have the same auditory inventory as English and Polish. In reality, of course, the inventory will be influenced by the rest of the phonological system of the language, because *tout se tient*.

Our findings are relevant for phonological theory, because they show that dispersion effects emerge automatically and non-teleologically in the phonology-phonetics interface, and that therefore phonological theory does not have to take them into account at higher levels such as in the phonological surface form (by dispersion constraints) or in the relation between underlying and surface form (by faithfulness constraints).

References

- Apoussidou, Diana (2006). On-line learning of underlying forms. *Rutgers Optimality Archive* **835**.
- Blevins, Juliette (2004). *Evolutionary phonology*. Oxford: Oxford University Press.
- Blutner, Reinhard (2001). Some aspects of optimality in natural language interpretation. *Journal of Semantics* **17**: 189–216.
- Boersma, Paul (1997). How we learn variation, optionality, and probability. *Proceedings of the Institute of Phonetic Sciences* (University of Amsterdam) **21**: 43–58.
- Boersma, Paul (1998). *Functional phonology: formalizing the interaction between articulatory and perceptual drives*. PhD thesis, University of Amsterdam. The Hague: Holland Academic Graphics.
- Boersma, Paul (2005). Some listener-oriented accounts of h-aspiré in French. *Rutgers Optimality Archive* **730**. To appear in *Lingua*.
- Boersma, Paul (2006). Prototypicality judgments as inverted perception. In Gisbert Fanselow, Caroline Féry, Matthias Schlesewsky & Ralf Vogel (eds.), *Gradedness in grammar*. Oxford: Oxford University Press. 167-184. [*Rutgers Optimality Archive* **742**, 2005]
- Boersma, Paul, & Paola Escudero (to appear). Learning to perceive a smaller L2 vowel inventory: an Optimality Theory account. To appear in a book edited by Peter Avery, Elan Dresher and Keren Rice. [*Rutgers Optimality Archive* **684**, 2004]
- Boersma, Paul, Paola Escudero, and Rachel Hayes (2003). Learning abstract phonological from auditory phonetic categories: An integrated model for the acquisition of language-specific sound categories. *Proceedings of the 15th International Congress of Phonetic Sciences*. 1013-1016.
- Boersma, Paul, and Bruce Hayes (2001). Empirical tests of the Gradual Learning Algorithm. *Linguistic Inquiry* **32**: 45–86.
- Boersma, Paul, and David Weenink (1992–2007). *Praat, a system for doing phonetics by computer*. Computer program, <http://www.praat.org>.
- Borgström, Carl H. (1940). *A linguistic survey of the Gaelic dialects of Scotland. Vol. I: The dialects of the outer Hebrides*. Norsk Tidsskrift for Sprogvidenskap, Suppl. Bind 1. Oslo: Aschehoug.
- Bradlow, Ann R. (1995). A comparative study of English and Spanish vowels. *Journal of the Acoustical Society of America* **97**: 1916–1924.
- Breen, Gavan, and Veronica Dobson (2005). Central Arrernte. *Journal of the International Phonetic Association* **35**: 249–254.
- Carlton, Terrance R. (1991). *Introduction to the phonological history of the Slavic languages*. Columbus: Slavica.
- Choi, John D. (1991). An acoustic study of Kabardian vowels. *Journal of the International Phonetic Association* **21(1)**: 4–12.

- Darwin, Charles (1859). *On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life*. London: John Murray.
- De Boer, Bart (1999). *Self-organisation in vowel systems*. PhD thesis, Vrije Universiteit Brussel.
- Denes, Peter (1955). Effect of duration on the perception of voicing. *Journal of the Acoustical Society of America* **27**: 761–764.
- Dupoux, Emmanuel, Kazuhiko Kakehi, Yuki Hirose, Christophe Pallier, Stanka Fitneva, and Jacques Mehler (1999). Epenthetic vowels in Japanese: a perceptual illusion. *Journal of Experimental Psychology: Human Perception and Performance* **25**: 1568–1578.
- Escudero, Paola, and Paul Boersma (2003). Modelling the perceptual development of phonological contrasts with Optimality Theory and the Gradual Learning Algorithm. In Sudha Arunachalam, Elsi Kaiser & Alexander Williams (eds.), *Proceedings of the 25th Annual Penn Linguistics Colloquium. Penn Working Papers in Linguistics* **8.1**: 71–85. [the correctly printed version is *Rutgers Optimality Archive* **439** (2001)]
- Escudero, Paola, and Paul Boersma (2004). Bridging the gap between L2 speech perception research and phonological theory. *Studies in Second Language Acquisition* **26**: 551–585.
- Fisher, Ronald A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London, Series A, Containing Papers of a Mathematical or Physical Character* **222**: 309–368.
- Flemming, Edward (1995). *Auditory representations in phonology*. Doctoral dissertation, UCLA. [Published 2002 by Routledge, New York & London]
- Flemming, Edward (2004). Contrast and perceptual distinctiveness. In Bruce Hayes, Robert Kirchner and Donca Steriade (eds.) *The phonetic bases of markedness*. Cambridge: Cambridge University Press. 232–276.
- Flemming, Edward (2005). Speech perception and phonological contrast. In David Pisoni and Robert Remez (eds.) *The handbook of speech perception*. Malden: Blackwell. 156–181.
- Flemming, Edward (2006). The role of distinctiveness constraints in phonology. Unpublished manuscript, MIT.
- Forrest, Karen, Gary Weismer, Paul Milenkovic, and Ronald Dougall (1988). Statistical analysis of word-initial voiceless obstruents: Preliminary data. *Journal of the Acoustical Society of America* **84**: 115–123.
- Gordon, Matthew, Paul Barthmaier, and Kathy Sands (2002). A cross-linguistic acoustic study of voiceless fricatives. *Journal of the International Phonetic Association* **32**: 141–174.
- Hamann, Silke (2004). Retroflex fricatives in Slavic languages. *Journal of the International Phonetic Association* **34**: 53–67.
- Harris, James W. (1969). *Spanish phonology*. Cambridge MA: MIT Press.
- Harrison, Phil (1997). The relative complexity of Catalan vowels and their perceptual correlates. *UCL Working Papers in Linguistics* **9**: 358–402.
- Hayes, Bruce, and Donca Steriade (2004). Introduction: the phonetic bases of phonological markedness. In Bruce Hayes, Robert Kirchner and Donca Steriade (eds.) *Phonetically based phonology*. Cambridge: Cambridge University Press. 1–33.
- Heffner, Roe-Merrill S. (1937). Notes on the length of vowels. *American Speech* **12**: 128–134.
- House, Arthur S., and Gordon H. Fairbanks (1953). The influence of consonant environment upon the secondary acoustical characteristics of vowels. *Journal of the Acoustical Society of America* **25**: 105–113.
- Jacquemot, Charlotte, Christophe Pallier, Denis LeBihan, Stanislas Dehaene, and Emmanuel Dupoux (2003). Phonological grammar shapes the auditory cortex: a functional Magnetic Resonance Imaging study. *Journal of Neuroscience* **23**: 9541–9546.
- Jäger, Gerhard (2003). Learning constraint sub-hierarchies: The Bidirectional Gradual Learning Algorithm. In Henk Zeevat & Reinhard Blutner (eds.) *Optimality Theory and pragmatics*. Basingstoke: Palgrave Macmillan. 251–287.
- Jakobson, Roman (1941). *Kindersprache, Aphasie und allgemeine Lautgesetze*. Uppsala.
- Jassem, Wiktor (2003). Polish. *Journal of the International Phonetic Association* **33**: 103–107.
- Johnson, Keith (1997). Speech perception without speaker normalization. In Keith Johnson & John W. Mullennix (eds.) *Talker variability in speech processing*. San Diego: Academic Press. 145–165.

- Johnson, Keith, Edward Flemming, and Richard Wright (1993). The hyperspace effect: phonetic targets are hyperarticulated. *Language* **69**: 505–528.
- Jones, Mark (2001). The historical development of retroflex fricatives in Polish: markedness, functionality, phonology and phonetics. Unpublished manuscript submitted for a Zdanowicz Prize for Polish Studies. Trinity College, Cambridge.
- Jongman, Allard, Ratreë Wayland, and Serena Wong (2000). Acoustic characteristics of English fricatives. *Journal of the Acoustical Society of America* **108**: 1252–1263.
- Kirchner, Robert (1998). *An effort-based approach to consonant lenition*. Doctoral dissertation, UCLA. [Published 2001 by Routledge, New York & London]
- Lacerda, Francisco (1997). Distributed memory representations generate the perceptual-magnet effect. Ms., Institute of Linguistics, Stockholm University.
- Ladefoged, Peter (2003). *Phonetic data analysis*. Malden, MA: Blackwell.
- Ladefoged, Peter, Jenny Ladefoged, Alice Turk, Kevin Hind, and St. John Skilton (1997). Phonetic structures of Scottish Gaelic. *UCLA Working Papers in Phonetics* **95**: 144–153.
- Ladefoged, Peter, and Zongji Wu (1984). Places of articulation: an investigation of Pekingese fricatives and affricates. *Journal of Phonetics* **12**: 267–278.
- Liljencrants, Johan, and Björn Lindblom (1972). Numerical simulations of vowel quality systems: the role of perceptual contrast. *Language* **48**: 839–862.
- Lindblom, Björn (1986). Phonetic universals in vowel systems. In John Ohala and Jeri Jaeger (eds.) *Experimental phonology*. Orlando: Academic Press. 13–44.
- Lindblom, Björn (1990). Explaining phonetic variation: a sketch of the H&H theory. In William Hardcastle and Alain Marchal (eds.) *Speech production and speech modelling*. Dordrecht: Kluwer. 403–439.
- Maddieson, Ian (1987). The Margi vowel system and labiodoricals. *Studies in African Linguistics* **18**: 327–355.
- McCarthy, John J. (2002). *A thematic guide to Optimality Theory*. Cambridge: Cambridge University Press.
- Mees, Inger, and Beverly Collins (1982). A phonetic description of the consonant system of Standard Dutch (ABN). *Journal of the International Phonetic Association* **12**: 2–12.
- Mielke, Jeffrey (2004). *The emergence of distinctive features*. PhD thesis, Ohio State University.
- Navarro Tomás, Tomás (1932). *Manual de pronunciación española*. 4th edition. Madrid: Centro de estudios históricos.
- Nowak, Paweł (2006). The role of vowel transitions and frication noise in the perception of Polish sibilants. *Journal of Phonetics* **43**: 139–152.
- Ohala, John (1981). The listener as a source of sound change. *Chicago Linguistic Society* **17**: 178–203.
- Oudeyer, Pierre-Yves (2006). *Self-organization in the evolution of speech*. Oxford University Press.
- Padgett, Jaye (2001). Contrast dispersion and Russian palatalization. In Elizabeth Hume and Keith Johnson (eds.), *The role of speech perception in phonology*. San Diego (CA): Academic Press. 187–218.
- Padgett, Jaye (2003a). The emergence of contrastive palatalization in Russian. In D. Eric Holt (ed.), *Optimality Theory and language change*. Dordrecht: Kluwer. 307–335.
- Padgett, Jaye (2003b). Contrast and post-velar fronting in Russian. *Natural Language and Linguistic Theory* **21**: 39–87.
- Padgett, Jaye (2004). Russian vowel reduction and Dispersion Theory. *Phonological studies* **7**: 81–96.
- Padgett, Jaye, and Marzena Zygis (2003). The evolution of sibilants in Polish and Russian. *ZAS Working Papers in Linguistics* **32**: 155–174.
- Pater, Joe (2004). Bridging the gap between receptive and productive development with minimally violable constraints. In René Kager, Joe Pater & Wim Zonneveld (eds.) *Constraints in phonological acquisition*. Cambridge: Cambridge University Press. 219–244.
- Pierrehumbert, Janet (2001). Exemplar dynamics: Word frequency, lenition, and contrast. In J. Bybee and P. Hopper (eds), *Frequency effects and the emergence of linguistic structure*. Amsterdam: John Benjamins. 137–157.
- Polivanov, Evgenij Dmitrievič (1931). La perception des sons d'une langue étrangère. *Travaux du Cercle Linguistique de Prague* **4**: 79–96. [English translation: The subjective nature of the

- perceptions of language sounds. In E.D. Polivanov (1974): *Selected works: articles on general linguistics*. The Hague: Mouton. 223–237]
- Prince, Alan, and Paul Smolensky (1993). *Optimality Theory: Constraint interaction in generative grammar*. [Published 2004 by Blackwell, London]
- Puppel, Stanisław, Jadwiga Nawrocka-Fisiak, and Halina Krassowska (1977). *A handbook of Polish pronunciation for English learners*. Warszawa: Państwowe Wydawnictwo Naukowe.
- Raphael, Lawrence J. (1972). Preceding vowel duration as a cue to the perception of the voicing characteristic of word-final consonants in American English. *Journal of the Acoustical Society of America* **51**: 1296–1303.
- Ringen, Catherine, and Pétur Helgason (2004). Distinctive [voice] does not imply regressive assimilation: evidence from Swedish. In Paul Boersma & Juan-Antonio Cutillas Espinosa (eds.), *Advances in Optimality Theory [International Journal of English Studies 4.2]*. 53–71.
- Rospond, Stanisław (1971). *Gramatyka historyczna języka polskiego*. Warszawa: Państwowe Wydawnictwo Naukowe.
- Sanders, Nathan (2003). *Opacity and sound change in the Polish lexicon*. Doctoral dissertation, University of California, Santa Cruz.
- Schwartz, Jean-Luc, Louis-Jean Boë, Nathalie Vallée, and Christian Abry (1997). The Dispersion-Focalization Theory of vowel systems. *Journal of Phonetics* **25**: 255–286.
- Shalev, Michael, Peter Ladefoged, and Peri Bhaskararao (1993). Phonetics of Toda. *UCLA Working Papers in Phonetics* **84**: 89–123. [also in *PILC Journal of Dravidic Studies* **4**: 21–56 (1994)]
- Smolensky, Paul (1996). On the comprehension/production dilemma in child language. *Linguistic Inquiry* **27**: 720–731.
- Stieber, Zdzisław (1952). *Rozwój fonologiczny języka polskiego*. Warszawa: Państwowe Wydawnictwo Naukowe. [Translated 1968 into English by E. Schwartz as *The phonological development of Polish*. Michigan Slavic Materials **8**. Ann Arbor: University of Michigan]
- Stone, Maureen, Alice Faber, Lawrence J. Raphael, and Tomas H. Shawker (1992). Cross-sectional tongue shape and linguopalatal contact patterns in [s], [ʃ], and [l]. *Journal of Phonetics* **20**: 253–270.
- Ten Bosch, Louis (1991). *On the structure of vowel systems: aspects of an extended vowel model using effort and contrast*. PhD thesis, Universiteit van Amsterdam.
- Tesar, Bruce (1995). *Computational Optimality Theory*. PhD thesis, University of Colorado.
- Tesar, Bruce (1997). An iterative strategy for learning metrical stress in Optimality Theory. In Elizabeth Hughes, Mary Hughes, and Annabel Greenhill (eds.), *Proceedings of the 21st Annual Boston University Conference on Language Development*. Somerville, Mass.: Cascadia. 615–626.
- Tesar, Bruce, and Paul Smolensky (1998). Learnability in Optimality Theory. *Linguistic Inquiry* **29**: 229–268.
- Tesar, Bruce, and Paul Smolensky (2000). *Learnability in Optimality Theory*. Cambridge, Mass.: MIT Press.
- Tingsabadh, Kalaya, and Arthur Abramson (1999). Thai. *Handbook of the International Phonetic Association*. Cambridge: Cambridge University Press. 147–150.
- Wedel, Andrew B. (2004). *Self-organization and categorical behavior in phonology*. PhD thesis, University of California at Santa Cruz.
- Wedel, Andrew B. (2006). Exemplar models, evolution and language change. *The Linguistic Review* **23**: 247–274.
- Zeevat, Henk, and Gerhard Jäger (2002). A reinterpretation of syntactic alignment. In D. de Jongh, M. Nilssenová, and H. Zeevat (eds.) *Proceedings of the Fourth International Tbilisi Symposium on Language, Logic and Computation*. University of Amsterdam.
- Zygis, Marzena (2003). The role of perception in Slavic sibilant systems. In Peter Kosta, Joanna Błaszczak, Jens Frasek, Ljudmila Geist, and Marzena Zygis (eds.) *Investigations into Formal Slavic Linguistics: Contributions of the Fourth European Conference on Formal Description of Slavic Languages*. Frankfurt: Peter Lang. 137–153.
- Zygis, Marzena, and Silke Hamann (2003). Perceptual and acoustic cues of Polish coronal fricatives. *Proceedings of the 15th International Conference of Phonetic Sciences*, Barcelona. 395–398.