THE UNIVERSITY OF CHICAGO

CUE SELECTION AND CATEGORY RESTRUCTURING IN SOUND CHANGE

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE HUMANITIES
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

DEPARTMENT OF LINGUISTICS

BY
JAMES P. KIRBY

CHICAGO, ILLINOIS
DECEMBER 2010

for rjm

CONTENTS

ACKNOWLEDGMENTS

A great many individuals have contributed in one way or another to the present work; I would like to take the time to thank them here.

The members of my dissertation committee, Alan Yu, John Goldsmith, Karen Livescu, Howard Nusbaum, and Jason Riggle, were extremely generous with their advice, perspectives, and time; this work is much richer for their input, and I owe a great intellectual debt to them all.

Completing this work in a timely fashion would not have been possible without access to a rich pool of empirical speech data collected and reported by other researchers. In particular, I would like to thank Natasha Warner, Allard Jongman, Joan Sereno, and Rachèl Kemps for granting me access to their production data on Dutch final obstruent devoicing.

My fellow graduate students at UCSD, especially Cindy Kilpatrick, Hannah Rohde, and Dan Brassil, provided friendship, solace, and advice in those first few soul-searching years and beyond. Conversations with my colleagues at the University of Chicago, especially Max Bane, Yaron McNabb, and Morgan Sonderegger, have shaped my thinking about the linguistic enterprise in general and many aspects of the present thesis in particular.

In the course of my academic career to date, I have been lucky to have the chance to study with a huge range of talented faculty. At UCSD, my MA advisors Farrell Ackermann, Sharon Rose, and Eric Baković, along with Andrew Kehler and Chris Barker, encouraged my nascent interests in applying computational techniques to linguistic problems, inspired me to continuous improvement, and graciously supported me even when I decided to switch departments. Jean Mark Gawron and Rob Malouf at SDSU taught me whatever I know about computational linguistics, and tried valiantly to impart much more than I no doubt retained. As Mark predicted, I will now forever remember where I was the first time the Viterbi algorithm was explained to me. As an undergraduate at the University of Wisconsin-Madison, Andrew Sihler, Joe Salmons, and Tom Purnell answered questions, wrote letters,

and encouraged my initial interest in the language sciences. Without their friendship and support, it is safe to say this document would not exist.

Finally, I would like to thank my family, especially my parents Monica and David, for love and encouragement, and to Rachel, my partner in this adventure and surely many more to come.

Chicago, November 2010

ABSTRACT

Changes to the realization of phonetic cues, such as vowel length or voice onset time, can have differential effects on the system of phonological categories. In some cases, variability or bias in phonetic realization may cause a contrast between categories to collapse, while in other cases, the contrast may persist through the phonologization of a redundant cue (Hyman, 1976). The goals of this dissertation are to better understand the subphonemic conditions under which a contrast is likely to survive and when it is likely to collapse, as well as to understand why certain cues are more likely to be phonologized than others.

I explore these questions by considering the transmission of speech sounds over a noisy channel (Shannon and Weaver, 1948), hypothesizing that when the precision of a contrast along one acoustic dimension is reduced, other dimensions may be enhanced to compensate (the PROBABILISTIC ENHANCEMENT HYPOTHESIS). Whether this results in phonologization or neutralization depends on both the degree to which the contrast is threatened as well as the informativeness of the cues that signal it.

In order to explore this hypothesis, phonological categories are modeled as finite mixtures, which provide a natural way to generate, classify, and cluster objects in a multivariate setting. These mixtures are then embedded in an agent-based simulation framework and used to simulate the ongoing process of phonologization of pitch in Seoul Korean (Silva, 2006a,b; Kang and Guion, 2008). The results demonstrate that adaptive enhancement can account for both cue selection as well as the appearance of cue trading in phonologization. Additional data from the incomplete neutralization of final voicing in Dutch (Warner, Jongman, Sereno and Kemps, 2004) are then used to show how variation in phonetic realization can influence the loss or maintenance of phonological categories. Together, these case studies illustrate how variation in production and perception of subphonemic cues can impact the system of phonological contrasts.

VITA

2000	B.A., Linguistics and Germanic Linguistics, University of Wisconsin-Madison
2005	M.A., Linguistics, University of California-San Diego
2007	M.A., Linguistics, University of Chicago
2010	Ph.D, Linguistics, University of Chicago

PUBLICATIONS

Kirby, James P. (to appear a). The role of probabilistic enhancement in phonologization. In A. Yu (ed.), *Origins of Sound Change: Approaches to Phonologization*. Oxford: Oxford University Press.

Kirby, James P. (to appear b). Acquisition of covert contrast: an unsupervised learning approach. In A. Baker, R. Baglini, T. Grinsell, J. Keane, and J. Thomas (eds.), *Proceedings from the Annual Meeting of the Chicago Linguistic Society 46*, Volume 2.

Riggle, Jason, Bane, Maximillian, Kirby, James, and Sylak, John (in press). Multilingual learning with parameter co-occurrence clustering. In *Proceedings of the North East Linguistics Society 39*.

Kirby, James P. (2010). Dialect experience in Vietnamese tone perception. *Journal of the Acoustical Society of America 127*(6):3749-3757.

Riggle, Jason, Bane, Maximiliian, King, Edward, Kirby, James, Rivers, Heather, Rosas, Evelyn, and Sylak, John (2007). Erculator: A Web application for constraint-based phonology. In M. Becker (ed.), *University of Massachusetts Occasional Papers in Linguistics 36: Papers in Theoretical and Computational Phonology*.

Kirby, James P. and Yu, Alan C. L. (2007). Lexical and phonotactic effects on word-likeness judgments in Cantonese. In *Proceedings of the XVI International Congress of the Phonetic Sciences*, 1389–1392.

Kirby, James P. (2006). The phonetics of Northern Vietnamese. In N. Duffield (ed.), *Vietnamese Online Grammar*, `http://www.vietnamese-grammar.group.shef.ac.uk`.

EDITED VOLUMES

Proceedings from the Annual Meeting of the Chicago Linguistic Society 43(1–2) (with M. Elliott, O. Sawada, E. Staraki, and S. Yoon). Chicago: Chicago Linguistic Society. 2007.

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION AND OVERVIEW

1.1 The role of phonetic variation in sound change

Linguists have long recognized variation in phonetic realization as a key component in sound change (Paul, 1889; Baudouin de Courtenay, 1895; Labov, Yaeger, and Steiner, 1972; Ohala, 1989). The multitude of phenomena known to be phonetic in one language but phonological in another, such as umlaut, palatalization, or nasalization, suggest that structural-phonological changes may arise via the reanalysis of phonetic variation. For instance, due to the timing of the necessary oral and velar articulations, vowels adjacent to nasals show partial nasalization in many languages (including English), conditioned by factors such as postnasal devoicing, postnasal frication, and vowel duration (Cohn, 1993; Beddor, 2009). Although these effects are allophonic or otherwise non-contrastive in some languages, in others, such as Old French and Northern Italian, there is evidence of the development of phonologically nasal vowels in VNC contexts to be conditioned by the same factors (Hajek, 1997; Sampson, 1999)

Based on the existence of a large number of similar phonetics–phonology 'doublets', linguists began to consider *how* a phonetic property might transition to a phonological one, a process Hyman (1972, 1973, 1976) dubbed PHONOLOGIZATION. Hyman famously illustrated the process using the example of tonogenesis, the process by which intrinsic fundamental frequency (F0) perturbations conditioned by the voicing specification of a prevocalic consonant come to be reinterpreted as inherent to the vowel and eventually as lexical tone (Matisoff, 1973; Hombert, Ohala, and Ewan, 1979). On Hyman's account of tonogenesis, sketched in Table 1.1, the universal[1], physiologically-based differences in vowel F0 (Stage I) first take

1. Note that the intrinsic F0 effect has also been demonstrated for languages with lexical tone contrasts such as Thai (Gandour, 1974, 1975), Yoruba (Hombert, 1975, 1977a), and Cantonese (Francis, Ciocca, Wong, and Chan, 2006); however, the perturbations persist for only a short period (10–30ms after vowel onset) as

on a language-specific form which is no longer strictly mechanical, i.e. they are to some extent under speaker control. At this point (Stage II), the pitch differences may be regarded as allophonic, conditioned by the initial consonant, but the stage has been set whereby a reanalysis may occur following the loss of the voicing contrast in the initials. If such a loss occurs, the syllabic contrast may be preserved via the PHONEMICIZATION of F0 (Stage III)[2].

Stage I	*Stage II*	*Stage III*
pá [‾—]	pá [‾—]	pá [‾—]
bá [‾⁄]	bă [‾⁄]	pă [‾⁄]

Table 1.1: Phonologization and phonemicization (after Hyman, 1976). Sparklines show the time course of F0 production for the vowel following the initial consonant.

An example of phonologization *in vivo* is provided by Seoul Korean, a language which maintains a three-way distinction between fortis, lenis, and aspirated stops in syllable-initial position (Table 1.2). While studies of Korean stop acoustics in the 1960s and 1970s found this contrast to be signaled largely by differences in voice onset time (VOT: Lisker and Abramson, 1964; Kim, 1965; Han and Weizman, 1970), subsequent studies have reported that lenis and aspirated stops are no longer distinguished by VOT in either production or perception, but rather that F0 has come to play a more central role (Kim et al., 2002; Silva, 2006b; Wright, 2007; Kang and Guion, 2008). In other words, the emergence of F0 as a primary cue to a category-level contrast in Korean may find its origins in what was originally intrinsic, mechanical, universal phonetic variation.

However, the Korean example illustrates two problems not addressed by the phonologization model as originally formulated. First, VOT and F0 are not the only cues relevant for the perception of this contrast: spectral tilt, the amplitude of the release burst, and the duration

opposed to languages like English, where differences may persist up to 100ms into the vowel (House and Fairbanks, 1953; Lehiste and Peterson, 1961).

2. Although Hyman's term for the Stage I > Stage II transition is reminiscent of *Phonologisierung* (Jakobson, 1931), Jakobson's use referred to the transition from an allophonic property to a phonemic one; thus (Hyman, 1976) suggests the term *phonemicization* for this subsequent transition.

manner		*1960s*	*2000s*	*gloss*
fortis	뿔	[p*ul]	[púl]	'horn'
lenis	불	[pul]	[p^{h}ùl]	'fire'
aspirated	풀	[p^{h}ul]	[p^{h}úl]	'grass'

Table 1.2: Phonologization of F0 in Seoul Korean.

of the stop closure have all been argued to play a role (Ahn, 1999; Cho et al., 2002; Kim et al., 2002; Wright, 2007). This is problem of cue SELECTION: determining *why* a certain cue is targeted in a given instance of phonologization, and not some other. In addition, as F0 has transitioned from a redundant cue to a primary cue, the previously primary cue, VOT, has been correspondingly less informative. This phenomenon – whereby phonologization of one cue is invariably accompanied by dephonologization of another – will be referred to as the problem of cue TRADING.

	voice contrast	*redundant effect*	*contrastive effect*	
English	/bæt/	[bæt]	[bæt]	'bat'
	/bæd/	[bæːt]	[bæːt]	'bad'
Dutch	/bat/	[bat]	[bat]	'benefit'
	/bad/	[bat]	[bat]	'bathe-1sg'

Table 1.3: The evolution of word-final obstruent devoicing in Dutch and English. In English, the redundant vowel length effect has been phonologized into a contrastive effect in this position, while in Dutch, the contrast has effectively been neutralized.

While examples like Seoul Korean are numerous, phonologization is not the only outcome of phonetic variation – merger of segments due to loss of phonetic contrast is of course a widespread phenomenon in sound change. In Dutch, for example, word-final obstruent devoicing has arguably resulted in homophony between word pairs such as *bat* 'benefit' and *bad* 'to bathe-1sg' (Lahiri, Schriefers, and Kuijpers, 1987). A similar process of word-final obstruent devoicing in English has not resulted in the loss of contrast in this position, however, since the

redundant effect of differences between the length of vowels preceding voiced and voiceless obstruents has been phonologized as the VOT differences were lost (Table 1.3).

While the English case once again illustrates the problems of SELECTION and TRADING, comparison with Dutch raises the additional problem of determining *whether* or not a contrast will be preserved or neutralized – the problem of category RESTRUCTURING.

This dissertation is concerned with explicating these three problems through careful examination of the empirical instances mentioned above. I propose that satisfying answers to these questions involves considering how both speaker and listener adapt to variation in their linguistic experience through optimizing the sometimes competing goals of communicative reliability and efficiency. In order to address the problems of SELECTION, TRADING, and RESTRUCTURING, a computationally explicit framework is described and tested using empirical data from the Korean and Dutch cases described above. The basic framework adopted is that of MIXTURE MODELS familiar from machine learning and statistical inference, which are used to model the production and perception of phonetic categories in a multivariate setting. In order to explore the influence of phonetic bias factors and cue reliability in sound change, agent-based simulations are used to model the interaction between members of a speech community. Finally, predictions about the loss or addition of category labels are made by way of computing the optimal trade-off between model fit and data coverage.

1.2 Sound change in three stages

Hyman's characterization of sound change separates the process into two distinct transitions: one by which universal, mechanical variation becomes language-specific, and a second stage at which this language-specific variation becomes contrastive (Table 1.4).

Subsequent research has tackled different issues raised by this model. One goal pursued by many researchers was to identify diagnostics that can be used to demarcate universal, mechanical, intrinsic phonetic variation (Stage I) from the language-specific, controlled, ex-

Stage I		Stage II		Stage III
universal phonetics	>	language-specific phonetics	>	phonology

Table 1.4: Three stages in sound change. After Hyman (2008).

trinsic variation (Stage II: Wang and Fillmore, 1961; Ladefoged, 1967; Ohala, 1981b; Solé, 1992, 1995, 2007), as well the language-specific phonetic variation (Stage II) from contrastive phonological variation (Stage III: Pierrehumbert, 1980; Kiparsky, 1995; Cohn, 1993; Hyman, 2008). Other researchers focused on identifying the set of PHONETIC PRECURSORS – articulatory, acoustic, and cognitive factors which constrain what is and isn't available as the input to phonologization (Hombert, 1977b; Ohala, 1981a, 1983, 1989, 1993a,b; Blevins, 2004; Moreton, 2002, 2008), with the goal of helping to define and delimit phonological typology (Ohala, 1989; Kiparsky, 1995; Blevins, 2004).

While many researchers focused on identifying the sources of phonetic variability that could (potentially) serve as input to phonologization, there remained the question of precisely *how* the process might unfold. Writing about the phonologization of F0, for instance, Ohala hypothesized

> [i]f these supposedly small fortuitous pitch contours following consonants can be used as perceptual cues by listeners, it is a small step beyond that to suppose that eventually these small pitch differences might be taken by listeners as the major acoustic cue differentiating the lexical items formerly differentiated by voicing or voice onset time. (Ohala, 1973: 10–11)

Although much of his research program has been devoted to the cataloging of likely phonetic precusors and their physiological underpinnings, Ohala took the additional step of developing a theory of how phonetic precursors, such as a 'small fortuitous pitch contour', might be reinterpreted by a listener. The core of Ohala's proposal is that sound change on an indi-

vidual level (a 'mini-sound change') is the result of listener misperception; in particular, of a listener's failure to take into account the effects of coarticulation or intrinsic variation (Ohala, 1981a, 1993b; see also Blevins, 2004, 2006). On such an account, the phonologization of a contextually conditioned feature such as F0 would arise due to a listener failing, for whatever reason, to perceptually compensate for the fact that the F0 perturbations at vowel onset are due to the presence of a neighboring consonant, instead reinterpreting them as a feature of the vowel itself (HYPOCORRECTION)[3]. On this view, phonologization is the result of innocent listener error, and the most common patterns of phonologization observed in the world's languages can be traced to universal physiological aspects of the human speech and hearing apparatus, a hypothesis that has received wide-spread empirical support (Ohala, 1981a, 1989, 1990, 1993a,b; Beddor, Krakow, and Goldstein, 1986; Hura, Lindblom, and Diehl, 1992; Guion, 1995; Plauché, Delogu, and Ohala, 1997; Hume and Johnson, 2001; Plauché, 2001; Beddor, Harnsberger, and Lindemann, 2002; Kavitskaya, 2002; Hayes, Kirchner, and Steriade, 2004; Przezdziecki, 2005).

1.2.1 The multivariate nature of speech

A key aspect of the speech signal highlighted by Ohala's theory is that of MULTIDIMENSIONALITY. Speech sound categories, be they phonemes or allophones, are not monolithic entities, bur rather are known to be identified on the basis of multiple acoustic-phonetic dimensions, which may serve as perceptual CUES to the categories (Delattre, Liberman, Cooper, and Gerstman, 1952; Liberman, Delattre, and Cooper, 1952; Cooper, 1953; Liberman, Cooper, Shankweiler, and Studdert-Kennedy, 1967; Liberman and Cooper, 1972).[4] Lisker (1978)

3. Similarly, Ohala argues that sound changes such as dissimilation are the result of listner HYPERCORRECTION, whereby a listener reinterprets the effects of intrinsic phonetic context as an extrinsic property of a sound segment.

4. Throughout this dissertation, the terms *cue* and *acoustic-phonetic dimension* will often be conflated. Strictly speaking, this is an abuse of terminology, since a given acoustic-phonetic dimension may not function as a perceptual cue for a given speaker, for a given contrast, or in a given language.

famously catalogued 16 possible cues to the perceptual distinction between English word-medial voiced and voiceless obstruents, including duration of the preceding vowel, F0 contour at vowel onset, and the timing of voice onset (VOT). While some cues are truly independent, others are often REDUNDANT, meaning that the value of one cue may be predicted on the basis of another. For example, in English, the F0 onset frequency of vowels is to some degree predictable from the VOT of the preceding consonant, with voiced obstruents (with short-lag VOT) having lower F0 than vowels following voiceless obstruents (with long-lag VOT). This is illustrated in Figure 1.1, which plots productions of /p/ and /b/ by speakers of American English. While the distinction between the categories is clear along the y axis (VOT), there is also some degree of separation along the x axis (F0).

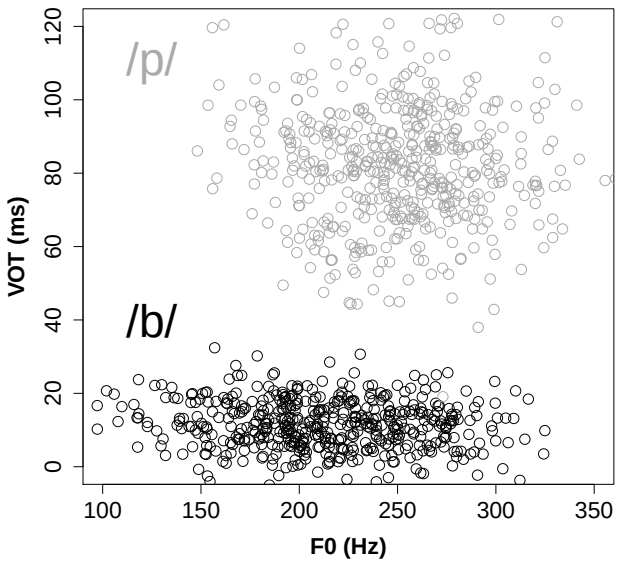

Figure 1.1: Redundancy between consonantal voice onset time (VOT) and fundamental frequency (F0) at vowel onset. Vowels following voiced obstruents (/b/) have lower F_0 than vowels following voiceless obstruents (/p/). Based on data from Clayards (2008).

While categories vary in the cues relevant for their perception, the contribution of each individual cue to the successful perception and identification of a given phonetic category can also vary, as a function of context, linguistic experience, and perceptual salience. For

example, while VOT is a highly robust indicator (or PRIMARY CUE) of the phonological [voice] feature of English obstruents in initial position, it is a poor indicator of that same feature in medial position (Clayards, 2008). Furthermore, in cases where a primary cue is for some reason uninformative – if it is masked by noise, for instance – attention can be re-focused on SECONDARY (redundant) cues (Repp, 1982). Attention to cue can be influenced by training and feedback (Francis, Baldwin, and Nusbaum, 2000), suggesting that the role of a given cue in signaling a contrast need not be static even over the lifetime of an individual, and may also be modulated by task, as demonstrated by Gordon, Eberhardt, and Rueckl (1993), who found that the perceptual role of canonically redundant F0 onset frequency increased when participants were distracted, while the attention to canonically primary VOT decreased.

1.3 Cue selection in phonologization

The multidimensional nature of the speech signal and the existence of multiple cues to phonetic contrasts are in some sense the empirical core of the phonologization framework. A question that naturally arises then is: how and why are *particular* cues phonologized in a given instance? This is the problem of cue SELECTION.

1.3.1 Psychophysical salience and linguistic experience

One possible answer to the SELECTION PROBLEM is that the likelihood of an acoustic dimension being targeted in phonologization is related to its degree of psychophysical salience. For instance, Holt and Lotto (2006) demonstrated that even when equally informative and psychophysically discriminable, listeners may still display a preference for categorizing non-speech stimuli based on one acoustic dimension over another. Francis et al. (2000) have shown that, prior to receiving instructive feedback, American English listeners relied more on formant transition cues than on frequency in the noise bursts preceding the transitions when categorizing obstruents, despite the fact that both cues reliably covaried with conso-

nant voicing. However, defining cue selection in terms of prelinguistic perceptual salience does not provide an explanation for the phonologization of those cues which are *a priori* less perceptually salient.

Another perceptual factor which may influence cue selection is linguistic experience, which has also been shown to exert considerable influence over the relative weight afforded cues in perception. For example, the length of a vowel preceding a medial or final obstruent is an important cue to the obstruent's phonological voicing specification in English (Port and Dalby, 1982; de Jong, 1991, 1995), but not in Arabic (Flege and Port, 1981) or Catalan (Solé, 2007). Examining this same cue, Nittrouer (2004) found that while native American English-speaking adults rely heavily on preceding vowel duration as a cue to the identity of a final consonant, children rely more heavily on dynamic properties of the signal, such as formant transitions, when making decisions about consonant identity. The relative importance of a cue may also vary by dialect: while Scottish English listeners rely heavily on vowel length to distinguish the /ɪ/ – /i/ contrast, Southern British English listeners pay more attention to formant frequency, facts which are also mirrored in their productions of these vowels (Escudero and Boersma, 2004). That the relative informativeness of acoustic-phonetic dimensions changes with linguistic experience has perhaps been most conclusively demonstrated by studies showing that by the first year of life, infants have reorganized their mapping from the acoustic input space to the categorical perceptual space in accordance with the input they receive (Werker and Tees, 1984; Kuhl, Williams, Lacerda, Stevens, and Lindblom, 1992; Kuhl, 2004; Kuhl, Stevens, Hayashi, Deguchi, Kiritani, and Iverson, 2006). There is also evidence that experience in non-linguistic domains can influence linguistic processing: when learning lexical tones, participants with musical training (who tend to give pitch movements greater perceptual weight) tend to outperform those without (Lee, Perrachione, Dees, and Wong, 2007; Wong and Perrachione, 2007; Chandrasekaran, Sampath, and Wong, 2010).

1.3.2 Cue selection and differential phonologization

The wide range of psychophysical and experiential effects is also reflected in the fact that different languages and dialects frequently diverge in their historical treatment of the same acoustic-phonetic dimensions. Table 1.5 shows the evolution of initial obstruent voicing in several dialects of Kammu (Khmu'), a Mon-Khmer language of Southeast Asia. Although F0 appears to have been phonologized to some degree in all three of the Kammu dialects shown, voice quality was phonologized one Western dialect, aspiration and F0 interact in another, while in the conservative Eastern dialect, the original voicing distinction has been preserved (Svantesson, 1983; Suwilai, 2001). Given that all the modern dialects developed from the same (Eastern) source, it is not at all obvious why the contrast should have evolved in the ways that it did. If the same phonetic precursors were available to all speakers, why did some Kammu speakers phonologize one cue and not another?

E. Kammu	*W. Kammu* (tone 1)	*W. Kammu* (tone 2)	*W. Kammu* (register)	*gloss*
buːc	pùːc	p^hùːc	pu̬c	'rice wine'
puːc	pûːc	p^húːc	pûc	'to take off clothes'
ɡlaːŋ	klàːŋ	k^hlàːŋ	kla̬ːŋ	'stone'
klaːŋ	klâːŋ	k^hláːŋ	klâːŋ	'eagle'

Table 1.5: Evolution of [voice] in Kammu. After Svantesson (1983); Suwilai (2001).

Kingston (1985, 2005, 2007) considers this question in his discussion of tonogenesis in the Athabaskan languages, which bears noteable similarities to the Kammu case. Kingston sought to explain the curious fact that among the tonal Athabaskan languages, which developed tones following the loss of final glottalic consonants, cognate forms have high tones in some languages, such as Slave and Chipewyan, but low tones in others, such as Gwichi'in and Navajo (Table 1.6).

Kingston's proposal hinges on two ideas, one focused on the listener and one on the speaker. The first is very much in line with Ohala's proposals and Blevins' (2004; 2006)

	Proto-Athabaskan	*Chipewyan* (High)	*Gwichi'in* (Low)	*Hupa* (non-tonal)
'smoke'	*ɬəd	ɬə̀r	ɬád	ɬid
'belly'	*wət'	bə́r	và̀d	mət'
'wife'	*ʔaˑd	ʔà	ʔád	ʔad
'scab'	*ɬuˑt'	ɬùr	ɬíd	ɬoh

Table 1.6: Representative Athabaskan cognate sets. t' = glottalic articulation, á = high tone, à = low tone, aˑ = full vowel, a = reduced vowel. (Examples from Krauss, 1979.)

notion of CHOICE: when multiple phonetic variants of a single phonological form are accurately perceived by the listener, the phonetic variant of the category prototype posited by the listener may differ from that intended by the speaker.[5] In the Athabaskan case, the idea would be that if a listener were to interpret the coarticulatory effect of the laryngeal articulation on F0 as an intended gesture, they would then encode this as part of the underlying specification for the contrast.[6]

The second part of Kingston's proposal differs from that of Ohala and Blevins in that it involves the speaker as well as the listener. In the Athabaskan case, Kingston proposes that speakers can and do actively manipulate phonetic cues to the laryngeal contrast in different ways. Building on earlier work by Krauss (1979) and Leer (1999), Kingston argues that, since the contraction of the cricothyroid and thyroaryetnoid muscles in the production of glottal consonants may occur independently (Kingston, 1985; Wright, Hargus, and Davis, 2002). If the glottal closure is affected only by contraction of the thyroarytenoid muscles, the outer vocal folds will remain slack and the voice quality of the adjacent vowel will be

5. The degree to which this type of reanalysis can be truly regarded as conscious choice on the part of the listener is not made clear by Blevins, Kingston, or Ohala.

6. While a plausible phonetically-based motivation for the differences in tonal evolution exists for Athabaskan, there are many well-known instances of similar tone reversals in the history of the Bantu languages, such as in Ruwund (Nash, 1994), Chiluba (van Spaandonck, 1971; Maddieson, 1976), and Tembo (Kaji, 1996), which are not so obviously amenable to such an account (Hyman, 2000). In addition, the realization of tone in many Bantu and other African tone systems is affected by a wide variety of phonological, morphological, and syntactic factors (see Kisseberth and Odden, 2003, for an overview).

breathy, accompanied by lowered F0. If, on the other hand, speakers simultaneously contract the cricothyroid muscle, the outer covers of the vocal folds will stretch, resulting in creaky phonation and heightened F0 on the adjacent vowel. The possibility of independent manipulation of these articulators allows Kingston to explain the fact that while some Athabaskan languages developed a high tone following the loss of laryngeal contrast, others developed a low tone, thereby providing a partial explanation of how cases of divergence in phonologization, such as seen in Kammu, might come about. Why speakers choose to exercise control over one cue versus another, however, remains an outstanding problem.

1.4 Category restructuring and licensing by cue

Up to this point, we have been discussing the problem of determining which cue might be phonologized, but Kingston also raises a slightly different problem: what determines whether or not a given cue will phonologize in the first place. Kingston notes that in other languages with phonological circumstances similar to Athabaskan, the loss of laryngeal contrast in stem-final position was *not* accompanied by a resulting phonologization, but simply resulted in positional neutralization. In Lithuanian, for example, a language which, like Proto-Athabaskan, once distinguished between voiced and voiceless obstruents in final position, the laryngeal contrast is now supported only before sonorants, as shown in Table 1.7; in other environments, such as word-finally, the obstruent voicing contrast has been neutralized.

form	*gloss*	*form*	*gloss*
*sil**p**nas*	'weak'	*sko**b**nis*	'table'
*daũ**g*** [dau**k**]	'much'	*kà**d*** [ka**t**]	'that'
*dè**g-t**i* [**kt**]	'burn-INF'	*míela**s** **d**raũgas* [**zd**]	'dear friend'

Table 1.7: Positional neutralization in Lithuanian.

While it involves a different cue (vowel length instead of vowel F0), the fates of the word-final laryngeal contrasts in Athabaskan/Lithuainian echo the Dutch/English scenario discussed at the beginning of this chapter. Both are instances of the RESTRUCTURING PROBLEM: why is phonologization the outcome in one language, but neutralization the outcome in another?

In discussing the Athabaskan/Lithuanian-type case, Kingston considers, and ultimately rejects, Steriade's LICENSING BY CUE proposal (Steriade, 1997) as a way of explaining the different outcomes. Sharing with the work of Ohala and Blevins the idea that common sound patterns find their source in universal properties of the human speech system, licensing by cue maintains that phonological contrasts are likely to be maintained in contexts where the acoustic cues to their identity are robust and easily perceived, and likely to be neutralized in contexts where they are reduced or otherwise obscured. Thus, the retention of laryngeal contrasts before sonorants in Lithuanian (Table 1.7) is due to the fact that the release and transition cues relevant for the [voice] contrast are robustly perceptible in that context; in word-final position and preceding nonsonorant obstruents, however, they are not, and neutralization is the result. Kingston points out that while licensing by cue might help to explain the fact that in Lithuanian, weak perceptual cues to word-final voicing resulted in a loss of contrast in this position, it sheds no light on the fact that some Athabaskan dialects, when faced with presumably a similar set of affairs, instead phonologized F0, because

> [i]f the phonetic correlates *available* to act as cues to a particular laryngeal contrast are the same in all languages where that contrast is found, then Lithuanian and Klamath speakers and listeners had at their disposal more or less the same materials to convey these contrasts...as Athabaskan speakers. Yet they failed to use them. The solution to this problem lies in the idea that speakers choose how they are going to pronounce a contrast, and therefore which of the available phonetic materials they're going to use. (2007:427)

Kingston conceives of phonetics not as 'something that happens to speakers', but something that can be actively manipulated to serve a communicative need (*ibid*). However, accepting that speakers can exert control over low-level phonetic details of the speech signal simply raises the SELECTION PROBLEM again – explaining how and why speakers wield this control in certain situations and not in others.

1.4.1 Dephonologization and contrast-driven enhancement

A slightly different approach to solving the SELECTION PROBLEM, found in earlier work by Kingston and colleagues, is based on the AUDITORY ENHANCEMENT HYPOTHESIS: the idea that cues are enhanced based on the degree to which they contribute to the perception of an INTEGRATED PERCEPTUAL PROPERTY, or IPP, which reinforces a phonological contrast (Diehl and Kluender, 1989; Kingston and Diehl, 1994; Diehl, 2008). In the case of the [voice] contrast, for example, cues with similar auditory properties, such as F1 and F0, are predicted to integrate, because both contribute to the amount of low-frequency energy present near a stop consonant. Cues such as closure duration and F0 would not be predicted to integrate precisely because they do not both contribute to such a property (Kingston, Diehl, Kirk, and Castleman, 2008). If cues are enhanced based on the degree to which they contribute to IPPs, this predicts that certain cues, such as closure duration, will not be enhanced, and thus presumably are less likely to phonologize. A similar view is put forth by Keyser and Stevens (2001, 2006), who argue that cues are targeted for enhancement as a means of reinforcing an existing phonological contrast (2001:287).

There are, however, some problems with the idea that phonologization is contingent on the presence of a contrastive phonological feature, such as [±voice]. First, there are cases where the phonologization of a feature is not dependent on its contrastiveness. In Punu, a Niger-Congo language spoken in Gabon, non-contrastive mid-vowel ATR harmony is phonologized out of what appears to be 'pure articulatory convenience' (Hyman, 2008),

with /ɛ/ > [e] and /ɔ/ > [o] before /i/ and /u/ (Kwenzi-Mikala, 1980). Second, there is the matter of the TRADING PROBLEM, the rather striking fact that in many instances, phonologization of one feature is accompanied by dephonologization of another:

> the phonologization process...must be interpreted literally: something becomes phonological, and *at the expense* of something else. (Hyman 1976:410)

This type of scenario is sometimes referred to as TRANSPHONOLOGIZATION (Hagège and Haudricourt, 1978; Hagège, 2004) or REPHONOLOGIZATION (Jakobson, 1931):

> une opposition ayant valeur distinctive est menacée de suppression; elle se maintient par déplacement d'un des deux termes, ou de l'opposition entière, un trait pertinent continuant, de tout manière, à distinguer ces termes [7] (Hagège and Haudricourt, 1978:75)

The TRADING PROBLEM is left unexplained by theories which aim to account for the SELECTION PROBLEM in terms of phonological contrast enhancement. In order to understand why phonologization is often accompanied by dephonologization, we need a theory of how cues are targeted for enhancement that takes into account both the functional aspects of linguistic communication as well as individual variation in linguistic experience.

1.4.2 Category restructuring and inhibited sound change

The RESTRUCTURING PROBLEM has sometimes been addressed in previous literature under the more general rubric of INHIBITED SOUND CHANGE. It has often been suggested that sound change is more likely to be inhibited when it would result in the neutralization of a lexically or morphologically informative contrast (Martinet, 1952; Campbell, 1996; Blevins and Garrett, 1998; Kingston, 2007; Blevins and Wedel, 2009; Silverman, 2010). One well-known example concerns the loss of final *–n* in Estonian (Anttila, 1989; Campbell, 1998).

7. "An opposition having distinctive value is threatened with suppression; it is maintained by displacement of one of the two terms, or the entire opposition, a relevant feature continuing, in any manner, to distinguish these terms" (my translation).

While final *–n* was lost throughout Estonian, the loss was inhibited in Northern Estonian dialects in just those cases when it would have led to homophony between verbal inflections.[8] In Southern Estonian dialects, this sound change took place across the board; it was presumably not inhibited in this same context because retention of *–ʔ* meant that the verbal forms could still be distinguished, as shown in Table 1.8.

Northern Estonian	*Southern Estonian*	*Proto-Balto-Finnic*	
kannan	kanna	*kanna-n	'I carry'
kanna	kannaʔ	*kanna-ʔ	'Carry!'

Table 1.8: Estonian verb forms after loss of *ʔ* and *n*. After Campbell (1998).

While the avoidance of what has been termed 'pernicious homophony' (Lass, 1980; Campbell, 1998; Blevins and Wedel, 2009) may well play a role in the inhibition of neutralization, it does not necessarily help to explain the RESTRUCTURING PROBLEM. In the Estonian case, for instance, the outcomes of interest are retention vs. loss of a morphological contrast, which is strictly speaking independent of transphonologization vs. loss of a phonological contrast. In any event, the RESTRUCTURING PROBLEM remains to be explained in those cases where homophony was *not* avoided, such as in the history of Mandarin Chinese, where historical sound changes resulted in a large number of previously distinct lexical items becoming homophonous (Baxter, 1992; Duanmu, 2000; Silverman, 2006). As illustrated in Table 1.9, there are no less than six morphemes pronounced *sù* in modern Mandarin, all but two of which may be reconstructed as having distinct pronunciations in Old Chinese. The existence of such cases suggests that homophony avoidance alone cannot explain why restructuring occurs in some instances but not in others.

8. In fact, the change did go through in certain parts of the paradigm, but homophony was avoided through other strategies such as consonant gradation and cliticization; see Campbell, 1998:90.

character	*gloss*	*Mandarin*	*Old Chinese*
粟	'millet'	sù	*sjok
肅	'solemn'	sù	*sjiwk
宿	'stay, lodge for the night'	sù	*sjuk
夙	'morning, early'	sù	*sjuk
素	'white'	sù	*saks
愬	'to complain, to tell'	sù	*sŋaks

Table 1.9: Homophonous morphemes in modern Mandarin and their Old Chinese reconstructions (following Pulleyblank, 1991; Baxter, 1992). After Rogers (2005).

1.5 An adaptive model of sound change

Solving the SELECTION, TRADING, and RESTRUCTURING problems simultaneously requires us to reconsider the roles of the speaker and listener in sound change, as well as the ways in which speakers exercise phonetic knowledge in the form of cue enhancement. Here, I take an functional approach to the problem, where the function of speech is assumed to be broadly communicative (cf. Liljencrants and Lindblom, 1972; Lindblom, 1990; Boersma, 1998; Flemming, 2001). By modeling changes to both the language-specific and structural aspects of sound system using a single mechanism, different scenarios which may lead to sound change may be explored and compared.

1.5.1 The noisy channel

As Jakobson famously remarked, '[w]e speak in order to be heard in order to be understood' (Jakobson, Fant, and Halle, 1951:13). This basic problem faced by language users finds a useful metaphor in the 'noisy channel' familiar from information theory (Shannon and Weaver, 1949). At one end of the channel is the speaker, who is attempting to send a message to the listener, the receiver at the other end. However, even under relatively ideal conditions, speech communication is fraught with difficulties, and a huge number of factors

– including, but by no means limited to, the influence of physiological, social, and cognitive constraints on speech production and perception – can introduce variability into the acoustic realization, potentially obscuring the speaker's intended message. In this work, asymmetries in speech production and perception, regardless of their ultimate source, will be collectively referred to as BIAS FACTORS (cf. Moreton, 2008; Garrett and Johnson, to appear). Setting aside for the moment questions about the source, nature, and influence of various bias factors, it is enough to simply note that many different types of bias can have a similar effect: the introduction of noise into the channel, much like interference on a telephone line (Figure 1.2).

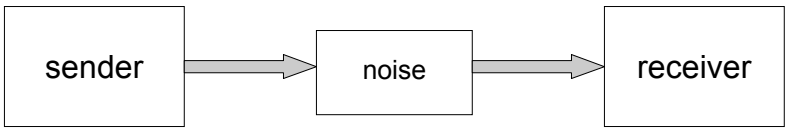

Figure 1.2: The noisy channel (after Shannon and Weaver, 1949).

To make this a bit more concrete, we may think of the speaker's (phonological) goal as being to transmit to the listener a sequence of labels, representing phonetic categories, each one signaled along multiple acoustic-phonetic dimensions. The listener's goal is to recover the speaker's intended sequence of labels based on the acoustic-phonetic information they receive[9]. All else being equal, the speaker's success is to some extent dependent on the PRECISION of the contrasts being transmitted – precision being determined based on the statistical distribution of acoustic-phonetic cues to the contrast in question. Precision may be reduced for a variety of reasons, including channel noise introduced by bias factors, or change in the system of contrast at the structural level, which may result in an increase or decrease in the number of categories competing over some acoustic-phonetic space. The

9. "Minimally, the talker needs to ensure that the linguistic units have sufficient discriminatory power for making the correct lexical identifications" (Lindblom, Guion, Hura, Moon, and Willerman, 1995:8). The present work ignores other potential sources of information such as phonotactic and syntactic context.

question of interest then becomes how language users respond in general to variation in the degree of contrast precision.

There is considerable evidence that, as listeners, language users are both aware of and able to adapt to the phonetic circumstances online. Remez, Rubin, Pisoni, and Carrell (1981) demonstrated that listeners can understand so-called 'sine-wave speech' by following the center frequencies of the first three formants, despite the overall reduction in available acoustic detail; similar results have been obtained for noise-vocoded normal speech (Shannon, Zeng, Kamath, Wygonski, and Ekelid, 1995; Davis, Johnsrude, Hervais-Adelman, Taylor, and McGettigan, 2005). Dupoux and Green (1997) showed that normal hearing listeners can, with some training, perform at close to normal levels of speech perception when exposed to severely time-compressed speech. And while the fact that telephone conversations take place in the 300 to 3000 Hz range might suggest that the most important information in the signal is contained in this bandwidth, listeners can categorize speech sounds with $> 90\%$ accuracy even when the signal is filtered to contain only frequencies below 800 Hz and above 4000 Hz (Lippmann, 1996).

There is also evidence that speakers adapt to changing communicative circumstances. An excellent example of this is provided by the study of Kang and Guion (2008), who show that the acoustics dimensions enhanced by speakers in production are related to those used to discriminate a contrast in perception. In particular, Korean speakers who distinguish voiced and voiceless stops on the basis of voice onset time tended to enhance that cue in clear speech, whereas speakers whose productions are distinguished more by F0 at the onset of a following vowel tended to enhance that cue. In conversational and citation-form contexts, neither grouped tended to enhance phonetic targets, suggesting both that degree of enhancement is (at least partly) a function of a speaker's assessment of the communicative context, and that the targets of phonetic enhancement can be modulated by a speaker's own experience. This particular example will form the basis of the simulations conducted in Chapter 4.

All of these studies provide support for the idea that language users are aware of, and able to compensate for, distortions in the signal. Assuming that speakers (i) have access to internal phonetic knowledge (Kingston and Diehl, 1994), (ii) equate the internal state of listeners with that of their own, and (iii) have some knowledge about the ways in which the communicative context (channel noise) might influence the precision of the contrast they are trying to transmit, speakers could exert phonetic control in an adaptive fashion by enhancing just those portions of the signal which would be most likely to ensure robust communication (Aylett and Turk, 2004; Diehl, 2008).

1.5.2 Probabilistic enhancement

This notion of ADAPTIVE ENHANCEMENT, whereby the speaker plays an active role in driving sound change, is reminiscent of the H(YPER)&H(YPO) THEORY of Lindblom (1990), in which the speaker is thought to keep a running estimate of the listener's need for information in the signal and adapts her productions accordingly, while balancing the speaker's own tacit preference for exerting the minimum articulatory effort necessary. So, while in general speakers may coarticulate as a means of reducing articulatory effort, they may HYPERARTICULATE in situations where the listener's informational needs are estimated to be high. The exact phonetic form produced in any given situation is thus co-determined by both these informational assumptions as well as articulatory constraints, such as vocal tract constraints (Ohala, 1981a, 1989), speaking rate (Moon and Lindblom, 1994) or a general preference for reduced articulatory effort (Bloomfield, 1933; Zipf, 1949; Lindblom, 1990; Kirchner, 1998; Flemming, 2001). In other words, the speaker must balance LISTENER-ORIENTED constraints ('be understood') with TALKER-ORIENTED constraints ('be efficient').

It must be noted that quantifying notions such as communicative efficiency or articulatory effort has proven notoriously difficult, and that experimental results do not always corroborate effort-based hypothesis. For example, Kaplan (2010) compared the degree of

intervocalic consonant lenition (voicing or spirantization) between normal talkers and intoxicated talkers, on the assumption that the latter group would tend to expend less articulatory effort. The results indicated that intoxicated subjects were more likely to produce less extreme articulations, but not that their productions were more or less lenited than those of sober talkers. The interpretation of these results for theories of sound change is complicated, in part because it is not clear if they should be taken to indicate that processes such as lenition are not driven by a preference for reduced articulatory effort (*contra* e.g. Kirchner, 1998), or if intoxicated speech is not an appropriate experimental condition under which to observe a reduction in such effort. Pouplier (2010) argues that careful or clear speech may be no more intrinsically effortful than casual speech, but that all speaking styles are equally optimal in their given contexts. These types of results suggest that articulatory economy, even if it can be adequately quantified, may not play a significant role in shaping the evolution of sound systems.

The basic finding that talkers will enhance certain aspects of the phonetic signal under certain conditions, however, is on firmer experimental ground. For example, Picheny, Durlach, and Braida (1986) found significant VOT lengthening for word-initial voiceless stops in English in clear speech conditions. Similarly, vowel intelligibility has been shown to improve under clear speech conditions in both English (Bradlow, 2002; Ferguson and Kewley-Port, 2002) and Spanish (Bradlow, 2002), for native as well as non-native listeners (Bradlow and Bent, 2002). These studies suggest that enhancement of phonetic targets is very real, even in the absence of a complete understanding of how enhancement might be modified by a talker's assessment of communicative efficiency or articulatory effort.

In arguing that talkers hyperarticulate by exaggerating phonetic targets in situations where intelligibility is degraded, H&H theory aims to explain differences in clear vs. conversational speech, but the same principle may also be applied as a motivating principle driving sound change (Lindblom et al., 1995). If the acoustic profiles of two speech sounds are

highly overlapping, listeners may have difficulty distinguishing between the two categories. If speakers were interested in improving intelligibility for the listener, they might choose to hyperarticulate in order to provide the listener with an acoustic target whose category label could be more accurately recovered. Exactly which cue(s) they choose to hyperarticulate will depend in large part on the degree to which a cue contributes to the successful perception and categorization of a phonetic contrast – what will be referred to as RELIABILITY (related to the notion of INFORMATIVENESS from Clayards, 2008).

The measure of experimental evidence suggests that it is unrealistic to assume that speakers are always optimal at assessing the communicative needs of listeners in a given context. However, the greater the potential loss in precision and the greater the potential increase in reliability, the more likely (or at least more motivated) a speaker may be to succeed. Evidence from research in speech perception suggests that the distribution of attention to individual cues may vary as the speech perception mechanism seeks out cues that are potentially more diagnostic under suboptimal listening conditions (Nusbaum and Schwab, 1986; Nusbaum and Magnuson, 1997; Magnuson and Nusbaum, 2007). As a first order approximation of these findings, this dissertation proposes a PROBABILISTIC model of adaptive enhancement which takes into account both contrast precision as well as cue reliability.

1.5.3 Cue restructuring and category restructuring

These same adaptive principles may be used to motivate a solution to the RESTRUCTURING PROBLEM as outlined above – determining the conditions under which the number and structure of phonetic category labels is likely to change. In this instance, the primary agent of change is argued to be the listener, as suggested by Ohala. Much as the speaker is thought to keep a running estimate of the informational needs of the listener, so too does the listener keep a running estimate of the efficiency of the set of category labels. As long as communication is sufficiently robust, the number of labels will be maintained, but the label

inventory will be reduced if it is not communicatively justified. In other words, the likelihood of neutralization is a function of channel noise and contrast precision (itself determined by the distributional properties of the acoustic-phonetic cues to the contrast). This process is modeled as ADAPTIVE REGULARIZATION, whereby the number and structure of phonetic categories is determined by optimally adjudicating between the model's fit to data as well as its overall complexity, measured by the number of parameters in the model. Since the parameterization grows with the number of components, this serves as a bias against large inventories of phonetic categories which are not justified by an increase in communicative accuracy.

1.6 Structure of the dissertation

In the preceding, I have identified several questions that arise when considering sound change from within a phonologization framework. Three challenges in particular stand out:

1. the SELECTION PROBLEM: predicting *which* cue will be the target in a phonologization process (or explaining why a given cue was targeted and not some other);

2. the RESTRUCTURING PROBLEM: predicting *when* a contrast will transphonologize and when it will neutralize (or explaining why a given contrast may be more or less apt to neutralize);

3. the TRADING PROBLEM: explaining *why* phonologization is so often accompanied by dephonologization.

I then proposed that formulating satisfying answers to these questions involves taking an adaptive perspective on the role of enhancement in sound change. The goal of this dissertation is to implement this model computationally, and illustrate its efficacy using real linguistic data. The basic computational framework adopted is that of MIXTURE MODELS

familiar from statistical machine learning, which are used to model the production and perception of phonetic categories in a multivariate setting. In order to explore the effects of different types of bias and cue reliability in sound change, these models are incorporated into computational agents in an agent-based simulation framework. Finally, the loss or addition of category labels is modeled by determining the optimal trade-off between model fit and data coverage.

The framework will then be used to explore two case studies: one illustrating the first transition in phonologization (the SELECTION and TRADING PROBLEMS, or factors influencing the reorganization of subphonemic cues) and one concerning the second (the RESTRUCTURING PROBLEM, or factors influencing the reorganization of category labels).

1.6.1 The mixture model of phonetic categories

Since the pioneering work at Haskins Labs in the 1950s (e.g. Cooper, Delattre, Liberman, Borst, and Gerstman, 1952), it is generally recognized that speech sound identification is influenced by a wide variety of highly variable acoustic dimensions, rather than being defined by a single set of invariant acoustic properties or distinctive features in the tradition of Jakobson (1931). Mixture models provide a convenient way of encoding this statistical variability while retaining a structural representation in the form of category labels. In addition, mixture models provide a natural means of relating speech production and speech perception (Lisker and Abramson, 1970; Nearey and Hogan, 1986; Pierrehumbert, 2002), provide a quantitative means of encoding the degree and extent of coarticulation (Recasens, Pallerès, and Fontdevila, 1997), and can easily model the effects of category frequency (Bybee, 2001). Chapter 2 begins by discussing two aspects of the speech signal crucial for understanding sound change, VARIABILITY and MULTIDIMENSIONALITY, that any representation should capture, then continues by laying out the mathematical foundations of mixture models, illustrating how they can be used to model aspects of speech production and percep-

tion, and discussing issues relevant to the classification and categorization of speech sound categories. Chapter 3 describes an agent-based iterated learning environment in which the mixture model is embedded in computational agents.

1.6.2 Phonologization as adaptive subphonemic restructuring

Chapter 4 explores the SELECTION and TRADING PROBLEMS by considering the case of phonologization of F0 in Seoul Korean in greater detail. While all studies of the Seoul Korean laryngeal contrast, from the 1960s to the present day, agree that some acoustic feature(s) other than VOT are crucial for distinguishing the Korean stop series, and while all have noted a role for F0, a number of other potentially relevant cues have also been observed, such as the duration of the stop closure, the amplitude of the burst, and the difference between the amplitude of the first two formants $H_1 - H_2$ (Cho et al., 2002; Kim et al., 2002; Wright, 2007). Indeed, the measurable presence of other potential cues has driven much of the heated debate in the literature regarding the proper phonological treatment of the Korean contrast (Kim, 1975; Jun, 1996; Kim, 2000; Ko, 2003; Kim and Duanmu, 2004; Silva, 2006a). While it appears that F0 already functioned as a redundant cue in the 1960s, it was not the only redundant cue. So why did it become the primary dimension along which the lenis/aspirated contrast was maintained? Chapter 4 explores several possible scenarios by combining the mixture modeling framework from Chapter 2 with the agent-based iterated learning environment described in Chapter 3 in order to simulate how the interaction of bias factors and adaptive enhancement effect changes in individuals' internal representations over time. The results show that both the SELECTION and TRADING PROBLEMS in Korean fall out of a model where adaptive subphonemic enhancement is itself a *response* to loss of contrast precision conditioned by phonetic bias factors.

1.6.3 Neutralization as adaptive category restructuring

Chapter 5 tackles the RESTRUCTURING PROBLEM by considering how individual variation in the production and perception of contrasts is reflected in a second type of sound change: a restructuring of phonetic category labels. The results of the simulations in Chapter 3 lend support to the idea that a language's inventory of category labels might be reduced in cases where a contrast has become particularly imprecise, i.e. when its communicative function is severely impaired. Chapter 5 explores this prediction using production data from both Korean as well as Dutch, a language in which near or total neutralization of a word-final voicing contrast is thought to obtain (Lahiri et al., 1987; Warner et al., 2004). While the contrast between Dutch final /t/ and /d/ can be quite difficult to perceive, it can be shown instrumentally and listeners show sensitivity to the contrast along several relevant cue dimensions (Warner et al., 2004), suggesting the possibility that the contrast may persist COVERTLY (Macken and Barton, 1980; Hewlett, 1988; Scobbie, Gibbon, Hardcastle, and Fletcher, 2000). This raises the question of just how imprecise a contrast would have to become before an adaptive listener would consider it neutralized. To explore this issue, Chapter 5 considers the Dutch case using data from Warner et al. (2004) in the context of a mixture model where the number of categories is not provided in advance, but instead determined based on a rational compromise between data coverage and model complexity, the BAYESIAN INFORMATION CRITERION (Schwarz, 1978). The results indicate that while neutralization is often predicted in cases of low contrast precision, multiple cue dimensions can interact with one another in such a way that contrasts which may be indiscriminable along one or more single dimensions may actually be separable in a higher-dimensional space, and that considerable individual-level variation in phonetic category structure may obtain in a given population.

1.6.4 Implications for the phonetics-phonology interface

Finally, Chapter 5 concludes with a discussion of the implications of the simulation results and the adaptive model for the phonetics-phonology interface more generally. While the simulation results provide an existence proof for an adaptive approach to the role of speech production and perception in sound change, this is different from providing experimental evidence that humans behave in this particular fashion. Chapter 5 discusses the types of experimental evidence which could be brought to bear on this issue.

In addition, although the mixture modeling approach to phonologization involves tracking statistics about individual continuous acoustic cue distributions, it does not wholly abandon the notion of a strict phonetics-phonology divide, since it organizes continuous dimensions with a set of discrete category labels. This final chapter closes with some thoughts on the implications of this view for the acquisition and transmission of phonetic categories, the cues which signal them, and the representation of speech sounds.

CHAPTER 2

MODELING SPEECH PRODUCTION AND PERCEPTION

This chapter describes the mixture modeling approach to categorization, and illustrates how it can be integrated into an agent-based simulation architecture to explore the dynamics of speech production and perception. The goal is to provide a brief technical introduction to mixture models and show how they map onto the questions of interest in this dissertation.

2.1 Desiderata for a representation of speech sound categories

Before discussing the particulars of the formal model, it is useful to review the properties of the speech signal that a model should take into consideration in order to provide a useful approximation of the human speech production and perception systems. Two of the most important properties are the following:

1. VARIABILITY: the acoustic realization of speech categories are highly variable.
2. MULTIDIMENSIONALITY: speech sounds may be distinguished by multiple acoustic-phonetic dimensions simultaneously.

We will consider each of these properties in turn.

2.1.1 Variability

Speech scientists have long struggled with the presence of variability in the speech signal. Practically every measurable dimension of speech, acoustic as well as articulatory, has been shown to vary in its precise realization. For instance, acoustic correlates of the initial [voice] contrast among obstruents in American English are not produced with exactly the same values from utterance to utterance, but instead can take on a range of values, which may also be influenced by a host of factors such as the sex of the speaker and the voicing specification

of the following consonant. In addition, some cues display more variability than others, as illustrated by the distributions of four cues (voice onset time, following vowel duration, burst amplitude, and F0 at vowel onset) to the English word-initial /b/–/p/ contrast shown in Figure 2.1.

Some variability is clearly crucial to accurate perception of contrasts. As seen in Figure 2.1, voiced and voiceless word-initial consonants in English have rather different distributions for VOT; the distribution of vowel duration for both categories varies depending on the nature of the following consonant; and the sex of the speaker is reflected in the distribution of F0 at vowel onset. Similarly, variation in the realization of the first two formants (F1 and F2) provides a means of distinguishing between different vowel categories.

In addition to variation which distinguishes *between* categories, however, there is also considerable variability *within* a given category; and generally speaking, it is the within-category variability that has been considered problematic, because it is not immediately obvious how to determine what portion of the variability constitutes information relevant to the perception of the contrast, what amount constitutes information about the identity of the speaker, and what amount is should be discounted as noise. Furthermore, the relative contribution of each of these individual sources of variance to the overall acoustic variance of a given dimension can vary considerably with context. For example, the fundamental frequency (F0) at vowel onset in the syllable /ba/, which is relevant for the distinction between /ba/ and /pa/, also encodes information about the age and sex of the speaker (Klatt and Klatt, 1990). However, as shown in panel D of Figure 2.1, the F0 value can be ambiguous between voiced and voiceless consonants as well as between speaker gender, leading to confusion regarding the interpretation of a given F0 value. The question of interest then becomes, how might listeners manage to recover approximations of the true underlying distributions?

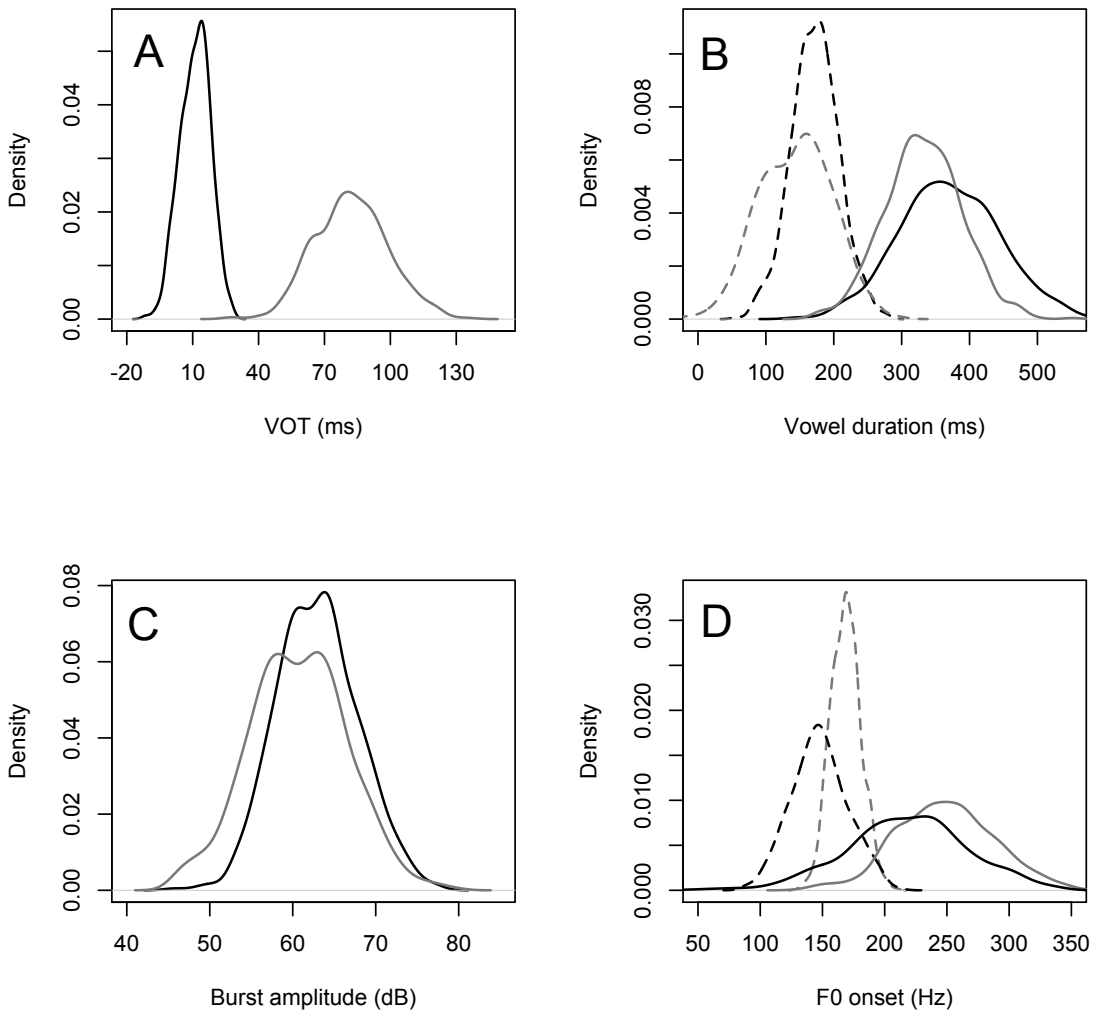

Figure 2.1: Kernel density plots of the distribution of cues to word-initial productions of /p/ and /b/ in American English. Black lines are instances of /b/, gray lines instances of /p/. A: voice onset time (VOT). B: vowel duration (solid lines represent voiced offsets, dashed lines voiceless offsets). C: burst amplitude. D: F0 at vowel onset (solid lines represent female speakers, dashed lines male speakers). Adapted from Clayards (2008).

Traditionally in speech perception, within-category acoustic variance was treated as a nuisance parameter obscuring the 'true', invariant acoustic properties which were thought to define speech sound categories. This is reflected in some of the early work on speech perception (e.g. Liberman, Harris, Hoffmann, and Griffith, 1957) in which all of the within-category variation in acoustic realization was considered noise to be stripped away by the perceptual system prior to speech sound categorization. However, it has since been increasingly recognized that listeners are extremely sensitive to within-category variability and use it to make decisions about phonetic category membership. Pisoni and Tash (1974) examined reaction times between acoustically identical and acoustically different stimuli belonging to the same phonetic category, and found that listeners responded significantly faster to acoustical identical instances of the same category. This same effect was replicated in an eye-tracking study by McMurray, Tanenhaus, and Aslin (2002), who found that within-category variation along a VOT continuum was mirrored in the length of fixations to cross-boundary lexical competitors, with fewer looks to alternatives for more prototypical exemplars. Miller and Volaitis (1989) demonstrated similar results using category goodness ratings. Thus instead of considering acoustic variance something to be removed from the signal, the focus has now shifted to better understanding how listeners make use of variability. This suggests that a model of speech sound categories should retain, rather than discard, information about acoustic variation within as well as between categories.

2.1.2 Multidimensionality

As reviewed in Chapter 1, speech sounds are not characterized simply by the presence or absence of a single invariant acoustic feature, but rather by multiple cues which are relevant (or potentially relevant) for accurate category perception. One reason categories vary across multiple dimensions is because a single articulatory gesture will often have multiple acoustic consequences. For example, the timing of the laryngeal gesture in initial stop consonants

relative to the supraglottal gesture can result in a variety of temporally overlapping acoustic effects, including the occurrence of glottal pulsing (voicing), aspiration, and F0 (pitch), and each of these may serve as a cue to the contrast between voiced and voiceless initials (House and Fairbanks, 1953; Lisker and Abramson, 1964).

A second reason why speech sound categories can vary across multiple acoustic-phonetic dimensions is because multiple gestures are involved in the production of a contrast. For example, the production of both voiced and voiceless bilabial stops in medial position in a language like English, such as in the words 'rabid' and 'rapid', involve a labial gesture to momentarily block the flow of air through the oral cavity, as well as vibration of the vocal folds. However, the relative timing of the events differs: in the case of the voiced stop, the vocal folds continue to vibrate throughout the oral closure, whereas in the case of the voiceless stop, glottal vibration is stopped (or at least reduced) during the oral closure. The timing relationship between these gestures is thus crucial to successful production of the contrast. In addition, as described above, a gesture may have multiple acoustic consequences: the rate of glottal vibration affects the F0 frequency going into and out of the closure, but also the length, while the precise nature of a labial gesture influences the amplitude of the release burst along with the length of the oral closure (Lisker, 1978).

Multidimensional variability is also accessible to listeners. Repp (1982) summarizes a number of studies indicating that listeners are sensitive to the multidimensional nature of the speech signal, and that they employ multiple sources of acoustic information when making decisions about the phonetic category of a stimulus. Much of this work, conducted at Haskins Labs in the 1940s and 1950s, concerns the study of what are often called TRADING RELATIONS. Repp defined a trading relation between two cues as

> "... a change in the setting of one cue (which, by itself, would have led to a change in the phonetic percept) can be offset by an opposed change in the setting of another cue so as to maintain the original phonetic percept." (Repp, 1982:87)

One example of a trading relation is that which exists between F2 at vowel onset and the amplitude of the release burst in the perception of the labial-alveolar distinction in CV syllables (Ohde and Stevens, 1983). Both high onset F2 and a high-amplitude release burst cue an alveolar response, while the opposite holds for the labial response. The trading relation that Ohde and Stevens demonstrated was that an *increase* in the value of onset F2 can be offset by a *decrease* in amplitude of the release burst. A second example comes Mann and Repp (1980), who demonstrated a trading relation between formant transitions and spectral center of gravity (COG) in the contrast between the English fricatives /s/ and /ʃ/. As COG was increased, so too did the proportion of /s/ responses, but this could be offset by splicing the formant transitions from /ʃ/ into the frication noise of /s/. The existence of this trading relation indicates that listeners are using both formant transitions as well as spectral COG in making judgements about the phonetic category of fricatives.

Repp (1982) made a further distinction between trading relations – how the interpretation of one cue varies based on the value of another relevant for the perception of the same contrast – and what he termed CONTEXT EFFECTS – how the interpretation of one cue is affected by preceding or following cues which are relevant for the perception of some other contrast. What is considered a cue to one contrast may be considered a context effect for another (Clayards, 2008). For example, the contrast between the vowels /u/ and /i/ is largely cued by the frequency of the second formant (F2), which is determined both by the degree of lip rounding as well as the position of the tongue in the oral cavity. Recall that F2 at vowel onset is also a cue to the /s/∼/ʃ/ contrast. If the formant transitions from a word like 'see' are spliced into white noise with a maximally ambiguous spectral COG, participants will tend to hear this as 'she' more often than they will hear 'shoe' if formant transitions from a word like 'sue' are spliced in, on account of the contextual effect of vowel rounding.

There is still some debate about whether certain acoustic events, such as the effect of vowel length on the category of a following vowel or the role of F2 in the perception of frica-

tive place of articulation, are best described as context effects or cues (Summerfield, 1975). Other researchers, such as Smits (1996), have instead sought to draw a distinction between AUDITORY CONTEXT (the information in the auditory time-frequency representation available at a given point in time) versus PHONOLOGICAL CONTEXT (the set of phonologically meaningful elements such as distinctive features, articulatory movements, segments, or syllables). For present purposes, it is not necessary to resolve this debate, but simply to note that, however they are described and delineated, phonologically relevant contrasts are signaled by a multitude of acoustic-phonetic features. However, it is worth noting the similarities that both trading relations and context effects bear to the phonologization processes described in Chapter 1. Recall that in Ohala's theory, one scenario that may lead to phonologization is when a listeners fails to take context effects (e.g. coarticulation) into account. In adaptive theories such as Lindblom's, such effects may be exacerbated by enhancement on the part of the speaker accommodating the listener's needs, reminiscent of the Repp's trading relations. Most importantly, trading relations mimic the TRADING PROBLEM in that increased attention to one cue is accompanied by decreased attention to another. The essential idea that will be developed here is that phonologization constitutes a sort of permanent cue trade through the reapportioning of relative cue reliability.

2.2 Classification, categorization, and clustering

The above discussion suggests that variability and multidimensionality are crucial aspects of the speech signal that should be taken into account by an adequate model of speech production and perception. The motivation here is not purely representational, but practical: we need to model both how listeners assign novel acoustic events to phonetic categories, as well as how they determine the set of phonetic category labels in the first place. In machine learning and statistical inference, the first problem – that of selecting the appropriate label for a new observation, given an extant set of category labels – is known as CLASSIFICATION (also

referred to as CATEGORIZATION in the psychological tradition), while the second problem – determining the structure of the categories themselves – is called CLUSTERING. The main difference between the two tasks is that in the case of categorization, the set of category labels is known (or assumed) from the outset, whereas in clustering, the number and structure of the category labels are inferred from the available data.

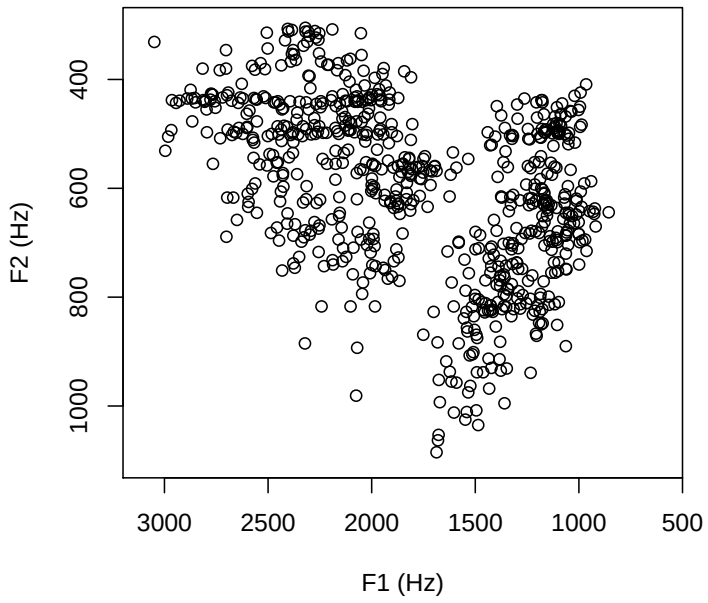

Figure 2.2: Raw vowel data. What is the underlying category structure?

To see the difference, consider the following example. Imagine you are given the vowel production data shown in Figure 2.2.[1] First, you might ask: are these observations all instances of a single vowel, or of multiple vowels? In clustering, the task is to determine the structure of the distribution(s) from which the observed data were drawn. Clustering is consider to be a type of UNSUPERVISED LEARNING, because the algorithm used to determine the number and structure of the underlying categories is given access only to the observations

1. These data are taken from Hillenbrand, Getty, Clark, and Wheeler (1995), available online at `http://homepages.wmich.edu/~hillenbr/voweldata.html`.

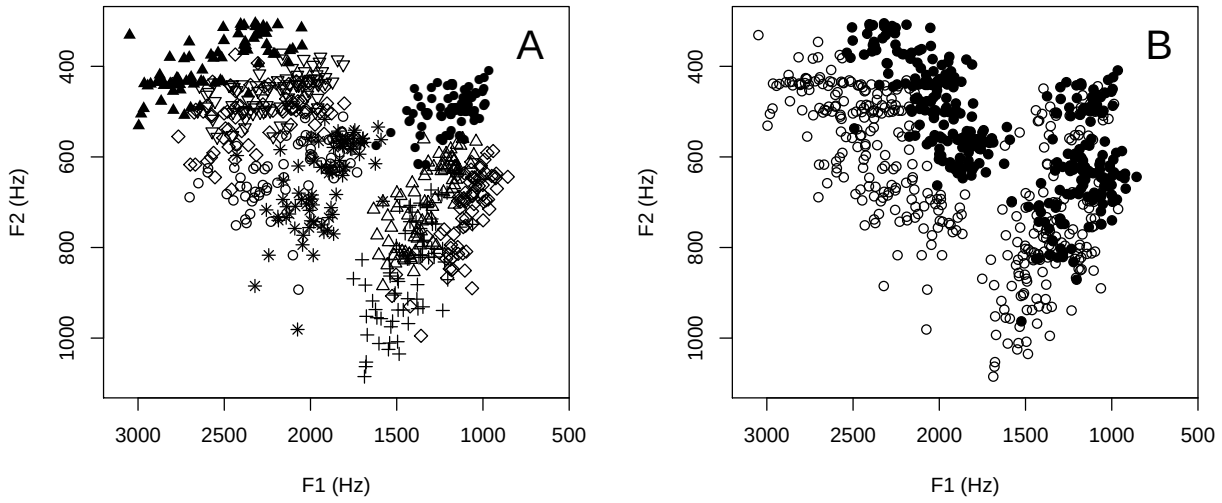

Figure 2.3: Two different possible clusterings/category structures for the Hillenbrand et al. vowel data. Panel A: classification based on 9 vowel categories. Panel B: classification based on two genders.

themselves, without prior knowledge of the category structure that generated them. Thus, in this example, you would not have access to any *a priori* knowledge of the underlying vowel category structure which may have generated the data in this figure.

On the other hand, you may already have some knowledge of the number and structure of the categories which generated the observations. For instance, you may know (or have some reason to believe) that the data in Figure 2.2 actually represent vowel productions of native American English speakers. Now, if you are given a new vowel token, you may want to assign it to a category based on previously observed data. In classification, the task is to assign new data to one of the pre-existing classes (or equivalently, to assign the new data point a class label) using a function known as a CLASSIFIER. Classification is considered a type of SUPERVISED LEARNING, because the classifier is trained on data where each data point consists of both an observation and its associated category label.

Classifiers can be built in a variety of ways, either using a clustering algorithm, or based on pre-existing, domain-specific knowledge. For example, depending on the purpose of the investigator, one may be interested in building a classifier to determine if new vowel measurements came from one of several vowels (Figure 2.3A) or from speakers of two different genders (Figure 2.3B); the form of the resulting classification function will vary accordingly.

In this and the following two chapters, we will be concerned chiefly with classification, under the assumption that the number of relevant speech sound categories is known in advance. In Chapter 5, we will return to the issue of clustering in greater detail, as it bears heavily on the RESTRUCTURING PROBLEM. First, however, we consider various theories of how human classification behavior might be formalized, and what empirical evidence exists that might help decide between them.

2.2.1 Prototypes, exemplars, and density estimation

One way to think about the categorization problem is in terms of typicality and generalization. For instance, in attempting to categorize a previously unencountered animal as, say, a bird or a non-bird, we might first develop a list of characteristics that we believe to be inherent to birds, whereby we may define the category BIRD by the set of features [+HAS-WINGS, +HAS-FEATHERS, +LAYS-EGGS, +FLIES, +HAS-A-BEAK], etc. However, problems with the idea of categories as defined by invariant properties begin to arise as soon as observations match in some, but not all, of the features. For instance, penguins lay eggs and have wings, beaks, and feathers, but cannot fly. Should penguins be classified as birds? If we decide that the answer is 'yes' even though penguins do not match on all the requisite bird-features, this seems to imply some sort of structure or hierarchy holds of the features, such that by dint of possessing certain features, an entity is a better, more canonical, or more typical exemplar of a category. Fleshing this out requires a theory of how typicality should be assessed.

Prototype models

One of the first answers proposed to the problem of typicality was to move to a different type of representation. In a PROTOTYPE MODEL (Posner and Keele, 1968; Reed, 1972; Rosch, 1973; Smith and Minda, 2000), each category is represented by a single prototype, usually defined as the most typical member of that category. In some versions of prototype theory, the prototype is a set of characteristic (binary) features, with new stimuli being assigned to a category based on the distance between the stimulus and the prototypes in terms of the number of shared features. For continuous data, the prototype is often defined as the average of all the members of the category; the probabilities with which a new data point is assigned a category label are calculated based on the (e.g. Euclidean) distance between each stimulus and the prototypes in the cue space.

Posner and Keele (1968) demonstrated prototype effect in categorization accuracy in a series of experiments showing that humans are able to more easily identify a prototype dot figure (e.g., a triangle) underlying a series of distorted patterns of dots. The more similar an exemplar to the category prototype, the more easily (accurately) it was classified (Figure 2.4). The similarity-choice model (Shepard, 1958; Luce, 1963) and the Fuzzy Logical Model of Perception (Oden and Massaro, 1978; Massaro and Oden, 1980; Massaro, 1987) are prototype models which have been applied to the problem of speech perception. Authors such as Kuhl (1991) have also shown that more prototypical members of a speech sound category are rated as better exemplars of that category than those which are less prototypical, and that typicality may exert a measurable influence on perceptual behavior.

Formally, prototype models may be defined either over a discrete (featural) or continuous space. In a discrete space, the prototype is a vector of e.g. the most frequent feature values, and the distance function will be something like edit or Hamming distance. In a continuous space such as the acoustic-phonetic space of speech, each category $k \in \{1, \ldots, K\}$ has a corresponding prototype μ_k, defined as the average of all members of that category; a

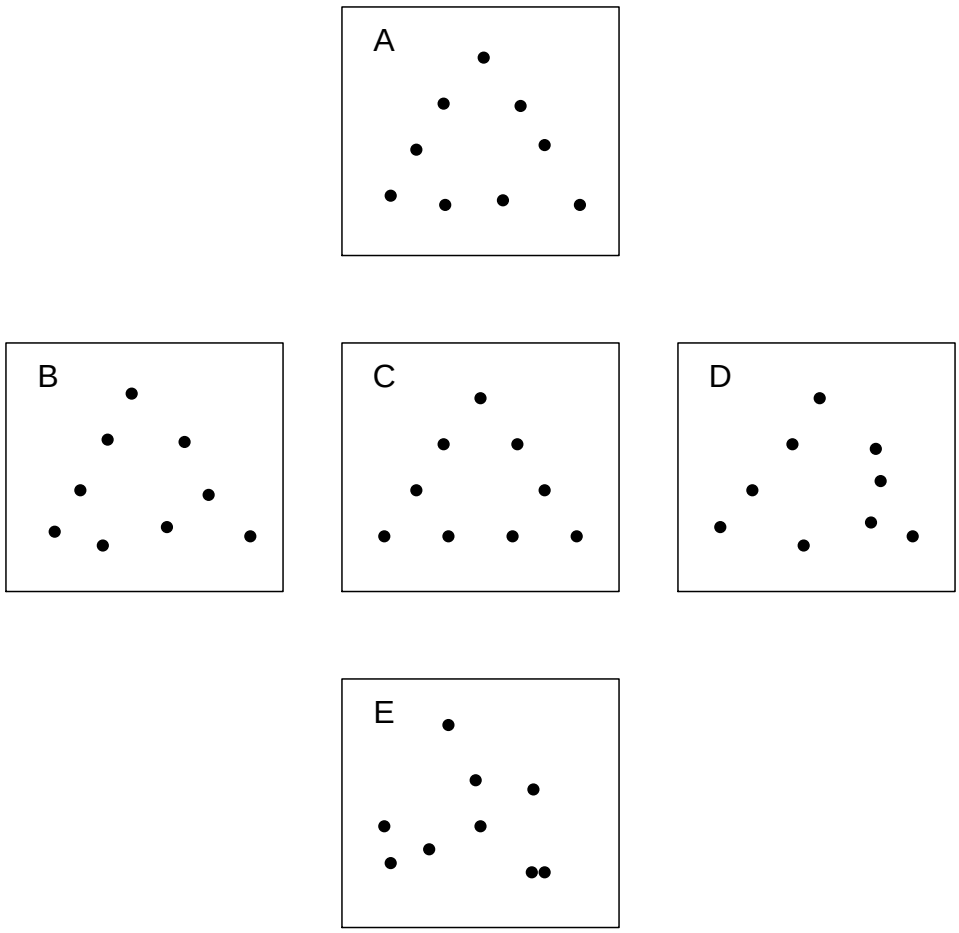

Figure 2.4: Categorization of dots (after Posner and Keele, 1968). Panel C represents the prototype; panels A, B, D, and E are increasingly distorted.

new stimulus x is assigned a category label that minimizes the distance (or maximizes the similarity) between x and the prototype. The distance function δ is often the Euclidean distance[2] between x and the prototype p in a hyperspace of S dimensions:

2. Other distance functions (e.g. Minkowski, Mahalanobis, quadratic, etc.) are also used; see Wang (2006).

$$\delta(x,p) = \left[\sum_{s=1}^{S}(x_s - p_s)^2\right]^{1/2} \tag{2.1}$$

While prototype theory accounted for some aspects of Posner and Keele's experiments, it failed to account for one extremely interesting one: the fact that, when asked to categorize a previously unseen arrangement of dots, participants were more accurate when that arrangement was close or identical to a previously seen arrangement, regardless of how far that arrangement was from the prototype. In other words, categorization accuracy appeared to be not just a function of typicality, but was based to some extent on previous experience as well. Furthermore, note that in prototype theories, only the *location* of the category prototypes in the cue/feature space is stored, not the measures of spread. This means that they do not encode information about within-category variability, one of the key aspects of the speech signal identified in §2.1.

Exemplar models

Both of these issues are addressed to some extent by EXEMPLAR MODELS (Nosofsky, 1986, 1990). In an exemplar model, categories are represented by sets of exemplars, represented as cloud of labelled points in cue space. The probability with which a stimulus is assigned a category label is based on a comparison with previously categorized exemplars across all categories, possibly modulated by a category bias. The Generalized Context Model (Nosofsky, 1986) is a well-known exemplar model defined for stimuli in psychological space; Lacerda (1995), Goldinger (1996), Johnson (1997), Pierrehumbert (2001), and Wedel (2004, 2006) define exemplar-theoretic models for word recognition.[3] The basic formalization of

3. As pointed out by Smits, Sereno, and Jongman (2006), there is no sublexical phoneme layer in many linguistic exemplar models such as those of Wedel (2004, 2006) or Pierrehumbert (2001). Instead, exemplars correspond to lexical items, which are stored with all of their fine-grained phonetic detail intact. However, there is no reason why an exemplar model could not be applied to the problem of identifying sublexical units (Smits et al., 2006); see also Kirchner, Moore, and Chen, in press.

categorization in an exemplar model is based on the LUCE-SHEPARD CHOICE RULE (Shepard, 1957, 1958; Luce, 1959). Given a set of n stored exemplars $y_1, y_2, \ldots, y_n$, and assuming for simplicity of exposition that there exist just two categories A and B, a novel stimulus x is assigned to category A with probability

$$P(A|x) = \frac{\beta_A \sum_{y \in A} \eta_{xy}}{\beta_A \sum_{y \in A} \eta_{xy} + \beta_B \sum_{y \in B} \eta_{xy}} \tag{2.2}$$

where η_{xy} is the similarity between exemplars x and y and β_A, β_B are pre-existing biases for categories A and B, respectively.[4]

One reason why exemplar models have been argued to be superior to prototype models is that they can produce prototypicality effects while predicting the 'old exemplar' advantage found by Posner and Keele. This is due the fact that, because old exemplars will tend to be near the category center, any new observation that is similar or identical to them will be predicted to have that category label with high probability. In addition, exemplar models predict a related effect that prototype models do not: even if an old exemplar is far from the prototype (category center), a novel observation that is close or identical to an old exemplar should be assigned the category label of that exemplar with high probability as well.

Density estimation and rational inference

One might assume that deciding between prototype and exemplar models of representation and categorization would be a strictly empirical matter, but the reality is somewhat more complex. While exemplar models tend to outperform prototype models in accounting for the variance in many visual categorization task results such as face recognition (Maddox and Ashby, 1998; Nosofsky, 1998), there is less empirical evidence available in the domain of

4. Some recent versions of exemplar theory, such as Nosofsky and Zaki (2002), use a deterministic, rather than a probabilistic decision rule. For more on decision rules, see §2.4.2 below.

speech. Inspired by experimental results which seemed to provide evidence for the categorical perception of speech sounds, researchers such as Liberman et al. (1957) modeled categories by learning a category boundary, which would then be employed in a deterministic fashion to label incoming acoustic stimuli. This approach is echoed in more recent work in detection theory (Macmillan and Creelman, 1981) and General Recognition Theory (Ashby and Perrin, 1988). Phonetic perception and classification have also been modeled using prototype theories (Samuel, 1982; Kuhl, 1991) and exemplar models (Pisoni, 1992; Nygaard and Pisoni, 1995). The superiority of the latter approach was argued for in part on the grounds that, as alluded to above, exemplar models can encode both recency and frequency effects observed in experimental studies of speech perception and word recognition.

In what may be the only study of its kind to date, Smits et al. (2006) explicitly considered the predictions of exemplar, prototype, and several related classes of models with respect to the classification of non-speech auditory stimuli. They noted that while different models vary in their underlying assumptions and therefore make fundamentally different claims about various aspects of categorization behavior, it is extremely difficult to distinguish between them experimentally, and the little work that has been done in this regard must be regarded as inconclusive. For instance, while Samuel (1982) takes the results of a phonetic categorization task using selective adaptation to support a prototype model of phonetic categories, his findings may also be explained by an exemplar model (a comparison Samuel did not consider at the time of the original experiment). Smits et al. (2006) conclude that the results of model fitting to their own experimental data are essentially ambiguous and are probably best explained by a hybrid model.

There are also theoretical reasons to be somewhat agnostic about making a hard choice between prototype and exemplar models of categorization. As noted by several authors (Estes, 1986; Ashby and Maddox, 1993; Ashby and Alfonso-Reese, 1995; Rosseel, 2002; Smits et al., 2006), both prototype and exemplar models have close analogs to statistical methods

of probability density estimation. In particular, many models of classification can be shown to be equivalent to an inductive process by which the observer estimates the likelihood that a novel stimulus x belongs to one of K categories $k = \{1, \ldots, K\}$ (Ashby and Alfonso-Reese, 1995). The probability that stimulus x is assigned category label k can then be found using Bayes' Rule:

$$P(k|x) = \frac{p(x|k)[(k)}{\sum_{k=1}^{K} p(x|k)p(k)} \tag{2.3}$$

A category label k may then be assigned either probabilistically (i.e. with probability $P(k|x)$) or deterministically (i.e such that $P(k|x)$ is maximized). This type of approach requires us to estimate some probability distributions: in particular, $p(k)$ (the prior probability of category k) as well as $p(x|k)$ (the probability of observing the stimulus if it is an exemplar of category k). On this view, the relevant distinction to be drawn is thus not between prototype vs. exemplar representations, but about how the relevant probabilities are best estimated.

It is useful at this juncture to define a distinction between PARAMETRIC and NONPARAMETRIC estimators. As used here, a parametric estimator is one which makes strong assumptions about the distribution of the observed data, while a nonparametric estimator makes no such assumptions. A classifier based on a parametric estimator will accordingly be referred to as a PARAMETRIC CLASSIFIER, while one based on a nonparametric estimator will be a NONPARAMETRIC CLASSIFIER (Ashby and Alfonso-Reese, 1995). On this definition, both prototype and decision bound models qualify as parametric classifiers, because they make strong assumptions about the structure of the category space (and as a result predict only linear or quadratic decision bounds). Exemplar models, on the other hand, are nonparametric, because they are not constrained by any underlying assumptions about category structure. Ashby and Alfonso-Reese (1995) show that exemplar models such as Nosofky's Generalized Context Model are equivalent to a classifier using a nonparametric (kernel) estimator (see also Rosseel, 2002).

Empirically speaking, there exists a considerable body of research (most of it involving the classification of visual stimuli) which argues for a strongly nonparametric model of human categorization behavior. McKinley and Nosofsky (1995) demonstrated that humans are not constrained to use linear or quadratic decision bounds when categorizing novel visual stimuli, and that categorization behavior was significantly better predicted by a nonparametric (exemplar) model than by a parametric (decision bound) one (see also Ashby and Waldron, 1999). The problems with taking such studies as evidence that human classification behavior is inherently nonparametric nonparametric are twofold. First, there are many empirical domains (such as speech) where the relevant distributions *do* appear to follow well-known parameterized distributions, such as the normal (Gaussian) distribution. Second, nonparametric approaches predict that, given enough training experience, human classification behavior should eventually come to resemble that of the underlying category structure, no matter how arbitrary. This is probably not the case, since human classification behavior clearly *is* limited, or at least preferentially constrained. For instance, McKinley and Nosofsky (1995) conducted a second experiment in which they had participants categorize visual stimuli for which the optimal likelihood classification boundary was both highly nonquadratic and not characterizable as a simple continuous curve. Even with continuous corrective feedback, only one-third of the participants in this experiment were able to exceed the classification accuracy of a quadratic (parametric) classifier, and another third were unable to even perform as well as a simple linear classifier. In a computational study of English and Japanese vowel category learning, Vallabha, McClelland, Pons, Werker, and Amano (2007) compared parametric and nonparametric versions of an online mixture estimation algorithm. They found the parametric algorithm significantly outperformed the nonparametric one, which they attribute to the nonparametric estimator's inherent lack of constraints on the underlying category structure.

These types of results suggests that, in its most extreme form, nonparametric density

estimation is likely too powerful a model of human classification behavior. This is especially true in the case of speech sounds, where the outstanding evidence reviewed above suggests that variability in the speech signal is accessible to and used by speakers and listeners. As the notion of category is ill-defined in exemplar-based approaches – the categories are, in a very real sense, defined by the experienced tokens themselves – it is not clear how variability in the signal should become information for the listener in such a model. Thus, in this dissertation, I shall restrict myself to a discussion of parametric estimators and classifiers. Should the balance of future empirical evidence come to favor nonparametric accounts, however, the computational framework discussed here can easily by modified to accommodate nonparametric estimators.

2.3 Finite mixture models

As suggested by the studies reviewed above, an accurate representation of speech sounds should encode both the variability and the multidimensional nature of the speech signal, while also enabling inference about the category-level (phonological) structure. One way to capture the distributional variance of a set of categories is with a FINITE MIXTURE MODEL, which models a statistical distribution as a weighted sum (or MIXTURE) of other distributions. Mixture models have a long history in mathematics and statistics, dating back to at least Pearson (1894). They are popular because they are powerful, flexible, and easy to implement, providing a natural way to represent the distribution and structure of a finite (although arbitrarily large) number of categories; provided with a sufficient number of components and accurate parameter estimates, finite mixture models can approximate any continuous density to arbitrary accuracy (Rosseel, 2002). The question of how the number of mixture components (categories) is selected is an important one, which will be considered in detail in Chapter 5. For now, we will assume that the number of categories which compete in a given region of multidimensional acoustic space, as well as the relevant dimensionality of

that space, are known in advance.

Mixture models, or some variation thereof, have been used extensively in previously work on modeling of speech sound categorization (Lisker and Abramson, 1970; Nearey and Hogan, 1986; Bybee, 2001; Pierrehumbert, 2001; Clayards, 2008; Feldman, Griffiths, and Morgan, 2009), even if they have not always been explicitly identified as such. More recently, mixture models have appeared under the guise of a MIXTURE OF GAUSSIANS (MOG) approach in work on the perceptual integration of multiple acoustic-phonetic dimensions (Toscano and McMurray, 2008; McMurray, Aslin, and Toscano, 2009; Toscano and McMurray, 2010) and the unsupervised induction of phonetic category structure (de Boer and Kuhl, 2003; Vallabha et al., 2007; see also Chapter 5).

2.3.1 Mixture models

A mixture model is always defined with reference to some set of data points, or OBSERVATIONS[5]. Let $\mathbf{x}$ be an observation represented as D-dimensional vector

$$\mathbf{x} = (x_1, \ldots, x_D) \in \mathbb{R}^D, \tag{2.4}$$

where each dimension is called a FEATURE. This featural representation is an abstraction of the observation; in the case of vowels, features might correspond to formants F1, F2, F3..., while in the case of consonants they might refer to things like voice onset time (VOT), burst amplitude, and so forth.

Now, assume we are provided with a finite number N of D-dimensional observations $\mathbf{X}$:

$$\mathbf{X} = \{\mathbf{x}_i; i = 1, \ldots, N\} \tag{2.5}$$

5. A note on notation: random (discrete or continuous) variables are capitalized (X) and individual values given in lowercase (x). When necessary, the position of a value in a sequence is indicated by a subscript (x_i). Vectors or matrices are shown in boldface ($\mathbf{x}$).

In a finite mixture model, we assume that these observations are sampled independently from an underlying distribution with a probability density function

$$f(\mathbf{x}; \boldsymbol{\pi}) = \sum_{k=1}^{K} \pi_k f_k(\mathbf{x}), \tag{2.6}$$

where π_k (the COMPONENT WEIGHT or MIXING COEFFICIENT) represents the probability that observation $\mathbf{x}_i$ belongs to the kth category (or COMPONENT) of the mixture with corresponding CLASS-CONDITIONAL DENSITY $f_k(\mathbf{x})$. K gives the total number of components in the mixture, which must obey the constraints that $0 \leq \pi_k \leq 1 \ \forall \ k \in \{1, \ldots, K\}$ and $\sum_{k=1}^{K} \pi_k = 1$.

2.3.2 Gaussian mixture models

Although nothing requires them to be so, in most cases the class-conditional density functions are assumed to be parametric, i.e. that the function $f(\mathbf{x})$ can be represented by a specific functional form containing a few adjustable parameters. The simplest and most widely used parametric mixture model by far assumes normal or GAUSSIAN density functions (Duda, Hart, and Stork, 2000; McLachlan and Peel, 2000). The (multivariate) Gaussian probability density function has the form

$$f(\mathbf{x}) = \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{2.7}$$

$$= \frac{1}{(2\pi)^{D/2} |\boldsymbol{\Sigma}|^{1/2}} \exp\{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})\}, \tag{2.8}$$

where the parameters are the D-dimensional MEAN VECTOR $\boldsymbol{\mu}$ and the $D \times D$-dimensional COVARIANCE MATRIX $\boldsymbol{\Sigma}$. The mean vector describes the central tendency of each of the feature dimensions; the covariance matrix contains the variance of each feature along the main diagonal, and the covariance between pairs of features of the other matrix positions.

We may now rewrite the mixture density function (Equation 2.6) as

$$f(\mathbf{x};\theta) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{2.9}$$

Equation 2.9 defines a GAUSSIAN MIXTURE MODEL (GMM), where $\mathbf{x}$ is a D-dimensional feature vector, π_k is the kth component weight, and $\theta = ((\pi_1, \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1), \ldots, (\pi_K, \boldsymbol{\mu}_K, \boldsymbol{\Sigma}_K))$ is a $K(D+2)$-parameter structure including the component weights as well as the mean vectors $\boldsymbol{\mu}_k$ and covariance matrices $\boldsymbol{\Sigma}_k$ of the D-variate component Gaussian densities $\mathcal{N}_1, \ldots, \mathcal{N}_K$.

Figure 2.5 gives the visualization of a univariate GMM with two components. The class-conditional densities (Equation 2.8) are shown in gray, while the mixture density (Equation 2.9) is outlined in black. Although more difficult to visualize, the mixture modeling approach extends straightforwardly to the multivariate case where $D = 2, 3, 4 \ldots$, as we would like it to for the representation of speech sounds. Exactly how each dimension contributes to the overall percept is the subject of ongoing investigation (see §2.4.3–2.4.4 below).

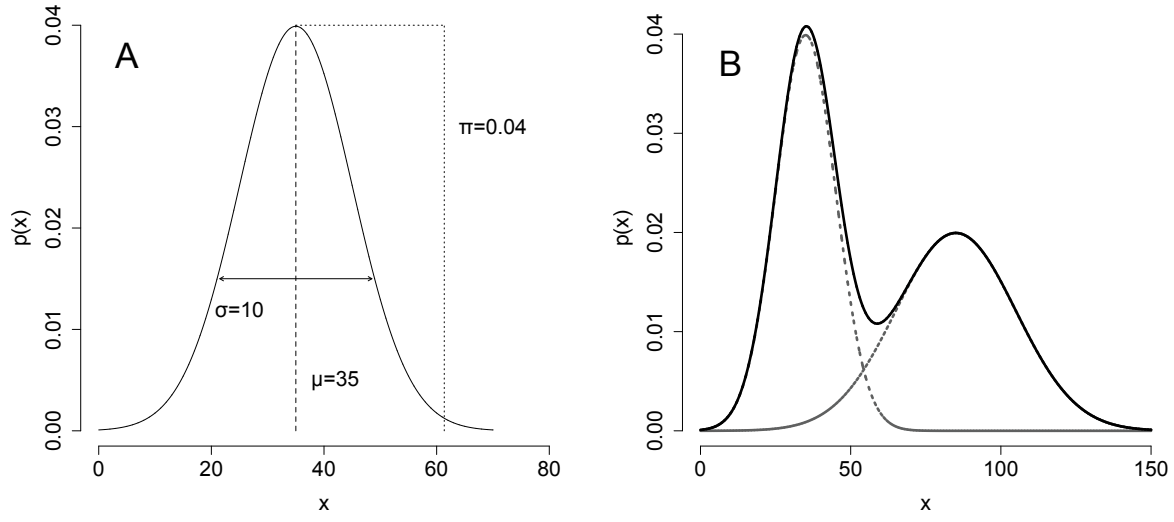

Figure 2.5: (A) Parameters of a Gaussian distribution for a single component (adapted from McMurray et al., 2009). (B) Two class-conditional Gaussians (dotted grey lines) and their mixture (solid black line).

2.3.3 Parameter estimation

In most statistical and machine learning problems of interest, the parameters θ of a GMM are not known in advance, but need to be estimated from the data. Traditionally, these parameters are computed using the expectation-maximizaton (EM) algorithm (Dempster, Laird, and Rubin, 1977) to obtain maximum likelihood (ML) or maximum *a posteriori* (MAP) parameter estimates, although other methods (such as Gibbs sampling) may also be used (see e.g. Bishop, 2006). Starting from an initial guess about θ, the EM algorithm alternates between computing a probability distribution over completions of missing data given the current model (the E-step) and then re-estimating the model parameters using these completions (the M-step). The E-step computes the conditional probability z_{ik} that observation $\mathbf{x}_i$ belongs to the kth component:

$$z_{ik} = \frac{\pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_j)} \tag{2.10}$$

In the M-step, the parameters θ are then re-estimated based on these conditional probabilities. This process continues until convergence, at which point the data may be partitioned into clusters or components on the basis of a decision rule (see §2.4.2 below). For more details on EM-based parameter estimation for multivariate Gaussian mixtures, see e.g. McLachlan and Peel (2000); Rosseel (2002); Fraley and Raftery (2002); Melnykov and Maitra (2010).

2.4 Modeling speech production and perception using GMMs

2.4.1 Modeling production: sampling from a density

One of the greatest conceptual and practical advantages of the GMM representation of phonetic categories is that it can be used to model both the production as well as the perception of phonetic categories, providing a formal and specific link between production data and categorization behavior (Nearey and Hogan, 1986; Solé, 2003). This allows a

listener's experience with instances of a phonetic category to form the basis for both the production of exemplars of that category as well as for determining the category label of novel exemplars.

The task of producing an instance of a phonetic category (be it a segment, subsegmental, or suprasegmental entity) is be modeled as sampling from a continuous multivariate probability density (technically, taking a point estimate from the approximate cumulative distribution function of the multivariate density as given in Equation 2.8) or by taking individual point estimates from a series of mixture of continuous univariate densities. Because the cue distributions are continuous, the true probability of any *given* value of x is in fact 0. However, we can define the probability of x falling into some *interval* of the cue space $[a, b]$:

$$P[a \leq X \leq b] = \int_a^b f(x)dx \tag{2.11}$$

for some arbitrarily small difference between a and b. In practice, the selection of a particular value is achieved by methods such as inversion sampling (where the cdf is equated to that of a pseudo random number generator) or rejection sampling; see Devroye (1986) for details.

2.4.2 Modeling perception: the ideal listener

The GMM for a given cue dimension d defines a probability density function $f(x_d)$; as illustrated above, sampling from this density may be used as a coarse approximation of the output of speech production. The task of the listener can be modeled as the reverse process: to determine the likelihood of a given category label c_k given an observation vector $\mathbf{x}$. If we consider the task of a listener to be choosing the speaker's most likely intended message given a set of cue values, and if we assume that the listener can make use of *all* of the information in the speech signal, possibly weighted in some way by its quality or reliability, we can construct a model of the behavior that would optimize this task. This is sometimes referred to as an IDEAL OBSERVER model, because the observer (here, the listener) is assumed

to behave optimally (ideally). Ideal observer models have been used to successfully model perception in a variety of domains and contexts including visual discrimination (Geisler, 1989), reading (Norris, 2006), word segmentation (Goldwater, 2006; Goldwater, Griffiths, and Johnson, 2009) and auditory word recognition (Flemming, 2007; Clayards et al., 2008).

An ideal observer model of speech perception works as follows. When attempting to identify the speaker's most likely intended message, the listener is assumed to have access to two sources of information: x (the information contained in the current speech signal) and $p(c_k)$ (their prior experience with the speech signal). Assuming that listeners track the statistical distributions of cues in speech (Maye, Werker, and Gerken, 2002; Clayards et al., 2008), they can use this information to estimate these probabilities. As discussed earlier, the probability that a given observation was generated by a mixture component may be found using Bayes rule. If we allow mixture component k to index phonetic category c_k, then Bayes rule may be equivalently used to estimate the probability of a category c_k given evidence x_d:

$$p(k|x_d) = \frac{p(x_d|c_k)p(c_k)}{\sum_{k=1}^{K} p(x_d|c_k)p(c_k)} \tag{2.12}$$

Equation 2.12 states that the *a posteriori* probability (APP) that the speaker uttered an instance of category c_k given the evidence that cue d takes on value x_d can be calculated given (i) the probability distribution for x_d given c_k and (ii) the prior probability of category c_k. The probabilities $p(x_d|c_k)$ are ML estimates based on the class-conditional Gaussian densities, and $p(c_k)$ may be estimated as the mixing coefficient π_k of component k.

Determining the likelihood that a given utterance x is an intended production of category c_k is logically distinct from the decision rule used by the listener to actually assign x a label. Nearey and Hogan (1986) discuss two possible decision rules, based on the APP of category membership as given by Equation 2.12. The first (deterministic) rule is

$$\text{Assign stimulus } x \text{ to category } c_k \text{ with the highest APP.} \tag{2.13}$$

while the second (probabilistic) rule is

Assign stimulus x to category c_k proportional to its relative strength of group membership. (2.14)

Nearey and Hogan take 'strength of group membership' to be defined over class-conditional Gaussian densities. To see the difference between the rules, consider a stimulus x which has probability 0.9 of belonging to category c_1 and probability 0.1 of belonging to category c_2 (as determined by Equation 2.12). Rule 2.14 will assign x the label c_1 90% of the time, and label c_2 10% of the time, while Rule 2.13 will always assign label c_1.

Nearey and Hogan (1986) attempted to decide between these rules on the basis of production data from Lisker and Abramson (1970). Both models fit the data well, with differences in goodness of fit too small to decide between them. In what follows, I will following previous research in assuming a probabilistic version of the decision rule on the grounds that it more accurately models human classification behavior (Shepard, 1957; Luce, 1959).

2.4.3 Cue independence and information integration

Of course, as discussed in great detail earlier, speech sound categories are defined by multiple cues. This raises the question of how cues are related to one another, and how listeners combine and integrate information from multiple cue dimensions. One assumption is that the distribution of each cue is conditionally independent from that of all others: knowing the value of one cue does not help in predicting the value of another cue.

Clayards (2008) shows how an ideal observer model may be extended to incorporate multiple, conditionally independent cues in word recognition, but the same principle may be applied to the categorization of subword units as well. The probability of a category c_k given a set of N D-dimensional observations $\{\mathbf{x}_1, \ldots, \mathbf{x}_D\}$ is the linear product of the probabilities of each individual cue, conditioned on the category c_k, normalized over all K

categories $c_1, \ldots, c_K$ (2.15):

$$P(k|\mathbf{x}_1, \ldots, \mathbf{x}_D) = \frac{p(x_1|k)p(x_2|k), \ldots, p(x_D|k)p(k)}{\sum_{i=1}^{K} p(x_1|k_i)p(x_2|k_i), \ldots, p(x_D|k_i)p(k_i)} \tag{2.15}$$

Note that from cue independence, it does not follow that the values of different cues fail to interact with one another in terms of determining categorization behavior. Figure 2.6 illustrates the interaction of two cues in the linear model. Panels A and B show the distribution of two cues d_1 and d_2, respectively, for two categories c_1 and c_2, which are assumed here to have equal prior probability. Cue d_1, having greater separation of means and less overlap than cue d_2, is the more informative cue (more on informativeness below). As a result, the listener's certainty about category membership, as indexed by the slope of the identification function in Panels C and D, is only slightly influenced by variation along dimension d_2, but extremely influenced by variation along dimension d_1.

Because this model is based on the assumption of conditional independence, it cannot capture relationships which can *only* be represented by multiple dimensions (Clayards, 2008; Goudbeek, Cutler, and Smits, 2008). There are, however, several empirical reasons to think that the strong independence assumption is a valid one. Clayards (2008) conducted a study of the conditional within-category independence of cues to the English stop contrast /p∼b/ in word-initial and word-medial position by calculating correlation coefficients between VOT, vowel duration, burst amplitude, voicing amplitude, and F0 onset/offset. In general, she found that correlation strength decreased with cue strength (as measured by d'). For word-medial stops, onset and offset F0 were rather more highly correlated (overall $R^2 = 0.64$), which is not unexpected given the relatively short time window involved. Several other minor correlations, such as burst amplitude and VOT ($R^2 = -0.22$), may have an articulatory basis, and did not seem to produce greater separation between the categories themselves. Although such correlations would need to be empirically established on a case-by-case basis,

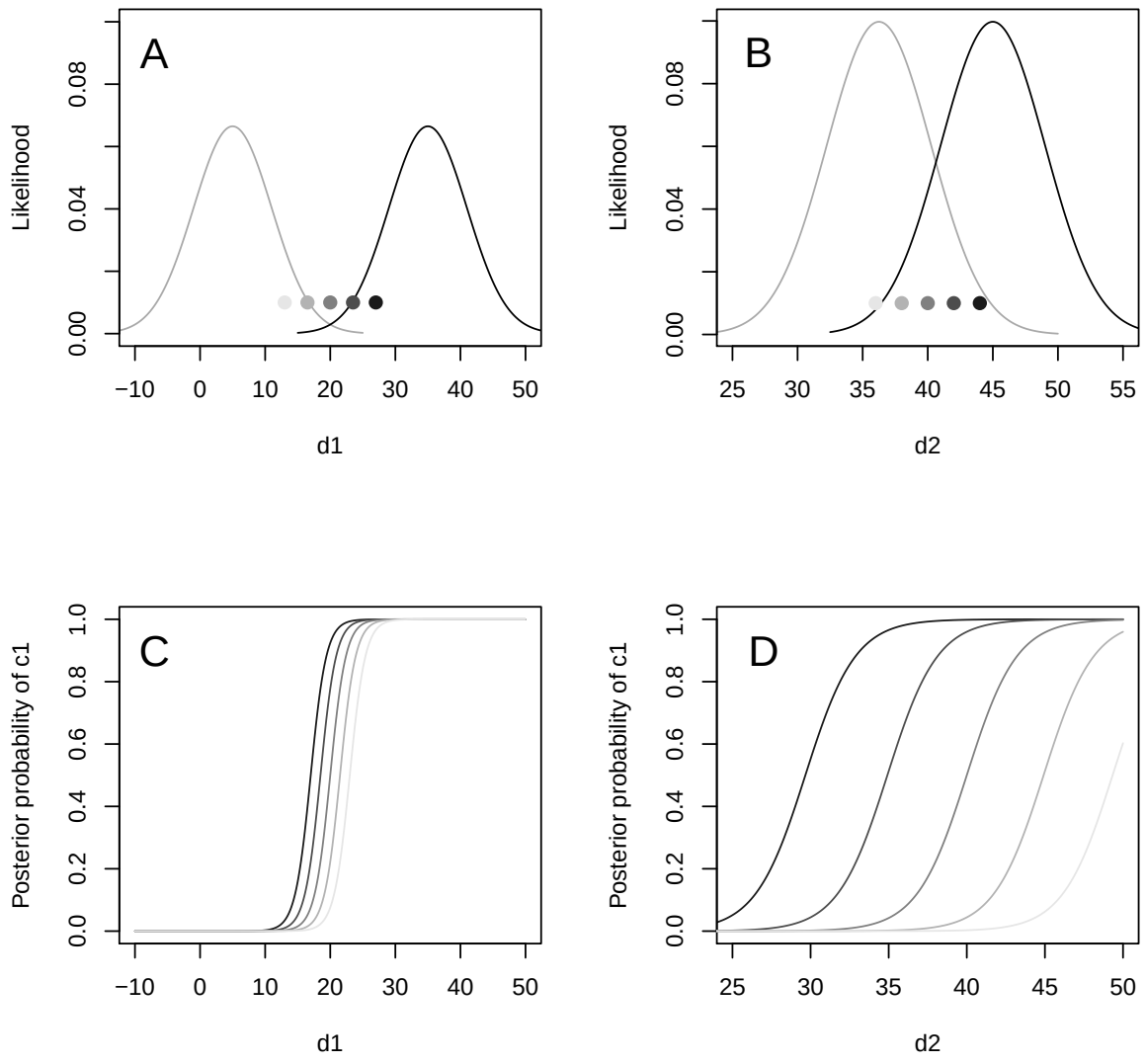

Figure 2.6: Hypothetical likelihood distributions illustrating how different cues combine in the linear model. Panel A: likelihood distribution of cue d_1 for categories c_1 (dark line) and c_2 (grey line). Panel B: likelihood distribution of cue d_2 for categories c_1 (dark line) and c_2 (grey line). Panel C: posterior probability of c_1 for all values of cue x and five values of y indicated by the shaded circles in Panel B. Panel D: posterior probability of c_1 for all values of cue d_1 and five values of d_2 indicated by the shaded circles in Panel A. Adapted from Clayards (2008).

the Clayards study is at least suggestive in this regard.[6]

More suggestive still are studies indicating a general preference for unidimensional solutions to categorization problems in human category learning, both in speech as well as in other domains. Goudbeek (2006) and Goudbeek et al. (2008) conducted a series of experiments studying the effects of distributional properties on the acquisition of novel (L2) phonetic categories in speech. Listeners were presented with novel vowel categories whose distributions varied along either one of two dimensions (duration and frequency) or both simultaneously. While listeners were clearly sensitive to the distributional properties of the input, they were better at acquiring and correctly identifying the novel contrast when it varied along only one dimension as opposed to both simultaneously. In particular, listeners had difficulty when the category structures to be learned were highly overlapping along individual dimensions, even when they were well-separated in higher-dimensional space. These results suggest a general preference for or sensitivity to unidimensional category solutions over multidimensional ones. This preference seems to apply not only to speech sound categorization, but to general stimulus categorization in humans as well. Ashby, Queller, and Berretty (1999) conducted a series of visual categorization experiments using lines that varied in either length, orientation, or both. In an unsupervised classification task (i.e., without feedback), learners were unable to learn nondimensional (arbitrary) category structures (although they were able to achieve near-optimal performance in a supervised learning scenario). Toscano and McMurray (2010) fit both a linear model of cue weighting and a true multivariate model to the same consonant categorization data, and found that the linear model provided a better fit. Thus a simple linear model may be sufficient for at least some categorization tasks, such as categorizing speech sounds.

6. Note that any cue with different means for two categories will be correlated with any other cue that has different means for the same two categories. These between-category differences can be captured by a model which assumes conditional independence; it is within-category correlations that are potentially problematic.

2.4.4 Cue reliability and cue weight

The ideal observer model predicts that listeners should make use of the probability distribution of *all* relevant cues when attempting to identify the speaker's intended utterance. The mere existence of multiple cues to phonetic categories does not, however, imply their equivalence: some cues provide more information about the perceptual identity of a sound than do others. This raises the question of how the WEIGHT afforded a given cue is to be determined. The weight of an acoustic-phonetic cue is a quantitative measure of how auditory information is integrated in perception (Holt and Lotto, 2006).

Several models, such as Nearey and colleagues' normal *a posteriori* probability (NAPP) model, estimate a cue's weight from its distributional statistics, i.e. purely as a function of its RELIABILITY in distinguishing between categories (Nearey and Hogan, 1986; Nearey, 1997; cf. Clayards et al., 2008; Toscano and McMurray, 2010). Reliability may be operationalized based on the degree to which the distributions of a cue overlap across categories: the less distributional overlap, the more reliable the cue; the more reliable the cue, the greater its role in determining the perceptual identity of an input. In other words, cue reliability is inversely proportional to cue variance (Clayards et al., 2008). Figure 2.7 illustrates this concept along a single cue dimension. The response curves in Figure 2.7B were computed using Equation 2.15 for each of the Gaussian mixtures in Figure 2.7A. Note that while the point at which the probability of categorizing stimulus x as belonging to either category c_1 or c_2 is equal (i.e., the point where the function crosses 0.5) is at the same point along the x-axis, the slope of the function is different, reflecting increased uncertainty in the case of the dashed distributions in Figure 2.7A.

One means of quantifying cue reliability is to use the detection-theoretic d' ('d-prime') statistic (Green and Swets, 1966; Ashby, 1992), a ratio of the difference in category means to the average variance:

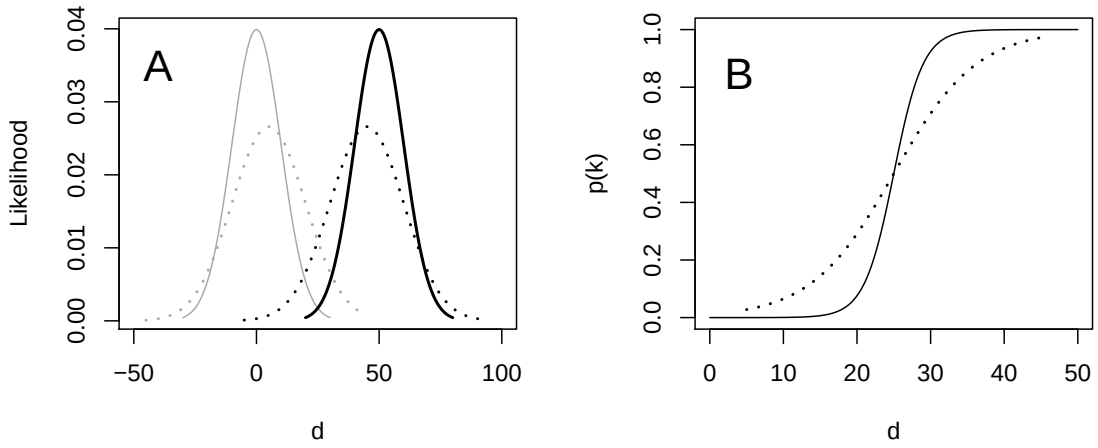

Figure 2.7: (A) Probability distributions of cue d for two categories c_1 (dark lines) and c_2 (light lines). Solid lines show a mixture where there is little overlap between the components, dashed lines a mixture with more overlap. (B) Optimal categorization functions given the distributions in (A). (Adapted from Clayards et al., 2008.)

$$d'(d) = \frac{(\mu_{d|k_1} - \mu_{d|k_2})^2}{(\sigma_{d|k_1} + \sigma_{d|k_2})/2} \tag{2.16}$$

By calculating a d' value for each cue $d \in \{1, \ldots, D\}$ and normalizin, each cue can be assigned a reliability value ω_d:

$$\omega_d = \frac{d'(d)}{\sum_{d=1}^{D} d'(d)} \tag{2.17}$$

Strictly speaking, however, a cue's weight is to some degree independent its reliability. Holt and Lotto (2006) discuss four types of factors which can influence the perceptual weight of a cue. The first two are distributional notions of INFORMATIVENESS (the degree of overlap between categories competing over some cue space) and VARIANCE (encompassing both the degree of within-category as well as between-category variance). Decreased overlap

between two categories along some cue dimension increases that dimension's informativeness, and hence reliability. The overall variance of a cue also has important implications for perceptual weight. While in general, large overall variance (independent of category) tends to be indicative of a highly informative perceptual dimension, large within-category variance can *decrease* reliability by contributing to increased overlap.

The distributional factors contributing to a cue's perceptual weight are to some extent captured by the d' statistic. However, there are other factors which can influence cue weighting which are not taken into account by d'. First, basic psychophysical considerations, such as the fact that discontinuities along some acoustic-phonetic dimensions induced a physical response of greater magnitude than others, may also exert an influence on the weight accorded a particular cue. Second, the informativeness of a cue is, as discussed extensively at the beginning of this chapter, inherently dynamic. As such, informativeness can also vary by task – a cue which is highly informative when attempted to identify a talker's gender or emotional state may be less informative when attempting to perform speech sound categorization or lexical decision, for instance. In the present work, we restrict ourselves to measuring cue reliability, mostly because it is much more obvious and straightforward and less contentious to operationalize. Thus in this model cue weights *per se* do not play an explicit role in speech perception (although cues are implicitly weighted in that their distributional variance impacts their role in assigning a category label). In general, then, we shall refer to cue reliability rather than cue weight, but the terms may sometimes be used interchangeably.

Other means of measuring cue weight have also been proposed. The method used here of calculating a normalized cue reliability statistic, bound to the interval [0,1], is similar to those of Holt and Lotto (2006) and Toscano and McMurray (2008, 2010)[7]. Clayards (2008) uses nonnormalized d' to measure cue weights. Escudero and Boersma (2004) and Morrison (2005)

7. Toscano and McMurray (2010) also contrast a cue-weighting model based on a d'-like statistic with a true multidimensional model in which the full covariance structure between cues is represented.

employ edge-based and logistic regression methods to determine relative cue weights based on listener responses to synthesized vowel continua. For present purposes, what is chiefly important is that (i) a cue's reliability may be inferred based on distributional properties of the acoustic input and (ii) reliability can be scaled in such a way as to be proportional to a probability. For more on modeling of cue weights and cue integration, see Holt and Lotto (2006); Toscano and McMurray (2010).

2.4.5 Classifier accuracy

Ashby and Maddox (1993) define the *optimal* classifier as one in which classification accuracy is maximized. Formally, this means that a given observation vector $\mathbf{x}$ is assigned to a category c_k for $k \in \{1, \ldots, K\}$ in such a way that $P(k|\mathbf{x})$ is maximized. Assuming that the *a priori* probabilities of the category indices $k = 1, \ldots, K$ and the class-conditional likelihoods $p(\mathbf{x}|k)$ are known (or can be estimated from the data), the posterior probability of each category index can be found using Bayes rule in the usual way:

$$p(k|x_1, \ldots, x_D) = \frac{p(x_1, \ldots, x_D|k)p(k)}{\sum_{k=1}^{K} p(x_1, \ldots, x_D|k)p(k)} \tag{2.18}$$

A deterministic classification rule assigns $\mathbf{x} = (x_1, \ldots, x_D)$ the category label $\hat{c}_k$ with the highest maximum *a posteriori* probability:

$$\hat{c}_k = \arg\max_{k \in K} p(k|x_1, \ldots, x_D)p(k) \tag{2.19}$$

This classifier is sometimes called a BAYES OPTIMAL CLASSIFIER (Duda et al., 2000). The error rate of this classifier may be expressed as

$$\epsilon = 1 - \sum_{k=1}^{K} \int p(\mathbf{x}|k)p(k)d\mathbf{x} \tag{2.20}$$

Figure 2.8 illustrates this behavior for the univariate case where $K = 2$ and both categories have equal prior probabilities.

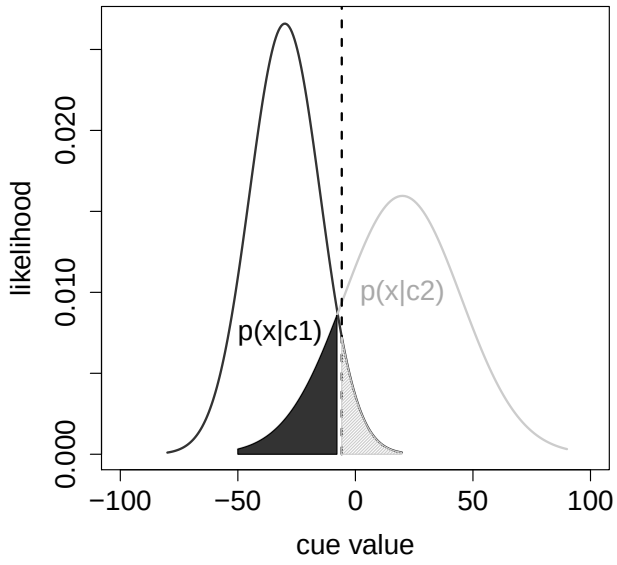

Figure 2.8: Bayes optimal decision boundary for two categories with equal prior probabilities. Light grey area shows the instances of c_1 that will be incorrectly labeled as c_2; dark grey area shows instances of c_2 that will be incorrectly labeled as c_1. Dashed line shows the optimal decision boundary. The total probability of error is calculated as the ratio of the shaded regions to the total region under both curves.

As noted by Ashby and Maddox (1993), optimal classifiers make strong assumptions and their predictions are not always in line with human classification behavior. However, they provides a useful benchmark against which to test other models, as the accuracy of a Bayes classifier provides a lower bound on the classification error that can be obtained for a given classification problem. As such, in this work, the error rate ϵ of a classifier with respect to a contrast will be defined as the error rate of the Bayes classifier. When reference is made to the 'error rate of a contrast', this is always restricted to mean between two or more phonetic categories that compete along the same number and type of acoustic-phonetic dimensions. Together with the reliability index ω, the error rate ϵ will figure centrally in the solution to the SELECTION PROBLEM as proposed and implemented in the following chapters.

2.5 Summary

This chapter has reviewed aspects of the speech signal that a model of speech production and perception should capture, including aspects which mirror the TRADING PROBLEM. Various theoretical formalizations of categorization behavior were considered. Parallels were drawn between problems in speech production/perception and in statistical categorization and category learning. The finite mixture modeling approach was advocated as an appropriate model of speech sound categories on the grounds that it captures both the empirically observed multidimensionality and variability of acoustic-phonetic cues. A fomal model of speech sound categorization was described and its application to speech perception illustrated. Finally, two important quantities were defined: the *reliability* ω of a cue dimension, defined as its normalized d' value, and the *error rate* ϵ of phonetic contrast, equivalent to that of a Bayes optimal classifier.

CHAPTER 3

AN AGENT-BASED SIMULATION ARCHITECTURE FOR MODELING SOUND CHANGE

In Chapter 2, I illustrated how speech sound categories may be usefully and to a certain approximation accurately modeled as finite mixtures, and discussed how these models may be used to simulate speech production and perception. But how does adopting this set of representational assumptions help deepen our understanding of sound change in general and phonologization in particular? In this chapter, I will lay out an agent-based simulation architecture in which the effects of arbitrary bias and probabilistic enhancement can be explored. The following chapter will then examine a specific empirical case in some detail, employing the computational framework to reason about how and why it changed in the way that it did, and not in some other way.

3.1 Simulating adaptive enhancement

Chapter 1 introduced the notion of ADAPTIVE ENHANCEMENT, the idea that speakers probabilistically enhance certain subphonemic aspects of their productions in response to their assessment of a listener's needs. In particular, it was suggested that phonetic bias factors which induce a loss of contrast precision may drive this type of enhancement. At a minimum, a computational test of this hypothesis must involve both a speaker, whose utterances are subject to some type of bias but who is cognizant of the listener's needs, and a listener, whose internal state is assumed by the speaker.[1] What is then needed is a framework in which arbitrary error-inducing biases can be introduced into a simulated communicative process, in order to observe the effects of probabilistic enhancement on the structure of the subphonemic cue space. By encoding prior beliefs and assumptions about enhancement and

1. In an extreme case, the speaker and listener could in fact be the same (Pierrehumbert, 2001).

bias into a computational model, we can uncover previously hidden implications of these assumptions. The study of sound change is particularly suited to computational simulation, since direct observation of subphonemic change in a population is a rather difficult and potentially labor-intensive process. Simulation allows for broad hypotheses to be tested prior to a more significant investment of research labor.

3.1.1 Computational models of sound change

Most computational models of sound change fall into one of two groups: ANALYTIC models, which focus on the mathematical properties of dynamic systems, and AGENT-BASED models, in which the internal state of populations change in the course of interactions between individuals. Analytic models, such as Niyogi and Berwick (1995, 1996), Wang, Ke, and Minett (2004), or Niyogi (2006), analyze sound changes in terms of dynamic systems, and derive equilibria (stable states) by considering a small number of parameters. Agent-based models, such as those of S. Kirby (1999), de Boer (2000, 2001), Wedel (2004, 2006), or de Boer and Zuidema (2009) are not so much an alternative as a complement to these analytic methods, in that they can be used to show how equilibria may arise spontaneously through the interaction of parameterized individuals (agents). Since the individual interactions depend stochastically on the current parameterization of the agents, the parameters are summarily stochastically modified by the outcome of the interactions.

Care must be exercised when evaluating the results of agent-based simulations, since the simple fact that the interaction of agents can result in behavior not obviously programmed into them is not particularly enlightening (Liberman, 2000). However, agent-based models can be used to derive more interesting results, such as the fact that, given certain constraints on the properties of the agents and their interactions, certain types of behavior emerge, but not others. It is in this spirit that an agent-based model is employed in the present work.

3.2 An agent-based model

To explore the temporal dynamics of the probabilistic enhancement hypothesis, an iterated learning environment (S. Kirby, 1999) was constructed in which ideal observer agents equipped with GMM models of phonetic categories could interact. Iterated learning differs from batch learning in that learning occurs after each input, making it both realistic and computationally tractable. In its simplest form, iterated learning consists of a single speaker-hearer (Pierrehumbert, 2001), although it can be easily extended to include multiple agents (Wedel, 2006; Baker, 2008) in order to address the dynamics of diffusion throughout a population (Niyogi, 2006; Sonderegger and Niyogi, 2010).

The simulation architecture presented here is employed to model a conversation between two agents, although the framework may be extended to accommodate arbitrary numbers of agents. Each agent is initialized as possessing a set of exemplar lists, one for each item in the lexicon; in the simulations discussed in §4.3, the lexicon consists of a minimal pair /pa ∼ p^ha/, although this can be increased without loss of generality. The contents of the exemplar lists at initialization will correspond to whether the agents are meant to represent adults or infants/children; here, the discussion is restricted to adult learners, who are presumed to have already acquired the sound system of the language and are aware of the number and distribution of the relevant speech sound categories. The length of the exemplar lists at initialization sets a bound on the agents' memories: as simulation time goes by, older, weaker exemplars will decay (be deleted from the lists) as newer, stronger exemplars are experienced (Pierrehumbert, 2001, 2002; Wedel, 2004, 2006).

3.2.1 Relations to exemplar theory

Because category membership in an ideal observer model is itself probabilistic, the simulation framework used here is closely related to exemplar-theoretic approaches to sound change (Johnson, 1997; Pierrehumbert, 2001; Wedel, 2004, 2006; Kirchner et al., in press; Garrett

and Johnson, to appear). In these models, computational agents do not literally store every token they encounter, but instead map experienced speech tokens onto a granular similarity space based on the token's similarity to a stored exemplar prototype; exemplars which fall between the cracks of this space are then encoded as identical (see also Kruschke, 1992). Each stored exemplar does not necessarily correspond to a unique perceptual experience per se, but rather to an 'equivalence class' of perceptual experiences (Pierrehumbert, 2001).

In many implementations of exemplar theory, stored exemplars are associated with a STRENGTH value, which decays over time. Exemplar strength may then be used in both speech production and perception: in production, exemplars of higher strength are more likely to be selected as production prototypes, while in perception, the probability with which tokens are mapped to stored exemplars is weighted by exemplar strength. In these implementations, the fundamental entity is the exemplar list, which again can differ from the sum total of experienced tokens depending on the granularity of the perceptual mapping.

The implementation described here differs slightly in that stored exemplars are not accessed directly in production or perception, but are instead used to estimate the parameters of the cue distributions relevant for some phonetic contrast. Similar to the implementations of Pierrehumbert and Wedel, experienced tokens are stored together with decay weights, but instead of directly influencing the selection of exemplars, decay weights are used to determine when an exemplar should be deleted from the list of tokens associated with a category label. Once the decay weight of a token falls below a certain threshold, it is deleted from the list and is no longer referenced during parameter estimation. When simulating speech production, values for each cue are simply sampled from each conditional density in the usual fashion. The advantage of this approach (due to the parametric assumptions made by adopting the GMM representation) is that it avoids several problems which arise when dealing with discrete lists of exemplars, such as implementating entrenchment or the addition of production noise in order to simulate generalization (Ashby and Maddox, 1993; Pierre-

humbert, 2001). However, it must be stressed that rigorous comparisons between previous implementations and the present have not been undertaken to ascertain how, if indeed at all, they differ in their empirical predictions; in the end, the approaches are grounded in very similar principles, and likely make very similar empirical predictions (see also Smits et al., 2006).

One important theoretical assumption made in the present implementation (and shared by many, but not all, other implementations) is that the *same* lists of experienced speech tokens are used by agents for both production *and* perception; to be more precise, each cue is represented by a single Gaussian mixture, the parameters of which are maximum likelihood estimates based on the current exemplar lists. This assumption that speech production and perception are based on the same set of phonetic exemplars is not universally shared; dual-route models of speech perception, in which e.g. word and phoneme recognition may take place using different, parallel mechanisms, have been argued for by phoneticians, psycholinguists, and neuroscientists (Liberman et al., 1967; Hickok and Poeppel, 2007; Norris and McQueen, 2008). Garrett and Johnson (to appear) extend this notion to exemplar-based models. Further research will be necessary to determine if models assuming separate exemplar lists more accurately reflect the empirical findings; the present framework could easily be modified to accommodate a dual-route architecture.

3.2.2 Conversing agents

Sound change in the agent-based model considered here is simulated as a bidirectional conversation between two agents, in which various aspects of the speech signal may be perturbed. Each round of the simulation is divided into the generation, modification, and classification of the production target. At each iteration, an agent acts either as a producer (generator) or as a perceiver (classifier). The roles are automatically reversed in the next iteration, so that each agent acts as producer and perceiver the same number of times in a given simulation.

The process repeats for some predetermined number of cycles, and can be interrupted at any time in order to examine the agents' current internal states. Agents are assumed to have equal sociolinguistic status (see Baker, 2008).

Initialization

Let K be a number of mixture components competing over a D-dimensional phonetic cue space. Each component $k \in \{1, \ldots, K\}$ indexes a unique member of a set C of phonetic category labels $C = \{c_1, \ldots, c_K\}$. A phonetic category c_k is associated with a list $\mathcal{E}_k = (e_1, \ldots, e_N)$, an N-length list of exemplars. An exemplar e_i is 4-tuple $(\mathbf{x}_i, c_k, t_i, \alpha_i)$, where

- $\mathbf{x}_i$ is D-dimensional column vector of phonetic cue values;
- c_k is a phonetic category label;
- t_i is the time at which e_i was added to $\mathcal{E}_k$;
- α_i is a dynamic memory weight, defined as

$$\alpha = \exp(-\frac{t_0 - t_i}{\tau}) \tag{3.1}$$

where t_0 is the current time, t_i is the time at which e_i was admitted to $\mathcal{E}_k$, and τ is a memory decay constant. In the simulations reported here, $\tau = 2000$ (Pierrehumbert, 2001).

To begin, K lists $\mathcal{E}_1, \ldots, \mathcal{E}_K$ are each seeded with N phonetic exemplars, with the values for each cue d being drawn from Gaussian distributions with parameters μ_d, σ_d:

$$\mathcal{N}(c_k|\mu_d, \sigma_d) = \frac{1}{(2\pi\sigma_d^2)^{\frac{1}{2}}} \exp\left\{-\frac{1}{2\sigma_d^2}(x - \mu_d)^2\right\} \tag{3.2}$$

Note that these are univariate Gaussian densities, a special case of the multivariate Gaussian densities described in Chapter 2.

Production

In the production phase of each iteration, the current talker selects a target category c_k based on the mixture weights π_k, themselves maximum likelihood estimates based on the length of the exemplar lists (which are equiprobable at initialization, although they generally do not remain so). Thus, for the case where $K = 2$, $p(c_1) = p(c_2) = 0.5$ at initialization; for the case where $K = 3$, $p(c_1) = p(c_2) = p(c_3) = 0.\bar{3}$ at initialization, etc. These probabilities may change throughout the course of the simulation as the frequency of the categories changes, e.g. due to the addition or loss of lexical items.

Once a target category c_k has been selected, a series of values $x_1, \ldots, x_D$ are sampled from the conditional densities $\mathcal{N}_d(x|k, \theta)$ (Equation 3.2) for all $d \in 1, \ldots, D$, where the parameters θ are maximum likelihood estimates based on $\mathcal{E}_k$. These values form an D-dimensional column vector $\mathbf{x}$, the PRODUCTION TARGET:

$$\mathbf{x} = (x_1, \ldots, x_D)^T \tag{3.3}$$

Enhancement

The next step in the generation of the production target involves determining whether or not to enhance some aspect of the utterance and, if so, precisely which aspect (*selection*). First, the talker estimates the listener's classification error rate ϵ for the target category based on the talker's own current exemplar lists, which are used as the basis for maximum likelihood estimates of the mean and variance of the GMMs for each cue. ϵ is defined as the error rate of a Bayes optimal classifier (cf. Figure 2.8 and §2.4.5):

$$\epsilon = 1 - \sum_{k=1}^{K} \int p(k)p(\mathbf{x}|k)d\mathbf{x} \tag{3.4}$$

The probability that any particular dimension d will be targeted for enhancement is an exponential function of ϵ and a BIAS CONSTANT $\beta \in [0, 1]$ which may be used to tune the system-wide importance or FUNCTIONAL LOAD of the contrast (Jakobson, 1941; Martinet, 1952; Hockett, 1955; Kingston, 2007). If β is set to 1, then the probability of enhancement is simply equal to ϵ. If the experimenter wishes to manipulate functional load, β can be reduced (making the probability of enhancement greater than ϵ) or increased (making the probability of enhancement less than ϵ).

$$P(\text{enhance}) = \epsilon^{\beta} \tag{3.5}$$

Thus, the likelihood of enhancement at any iteration is inversely proportional to the contrast precision (ϵ) scaled by the importance of the contrast (β). In the simulations reported here, β was arbitrarily fixed (not fit) at 0.5.

In the event that an utterance is selected for enhancement during a given iteration, the next step is to decide which cue in particular should be enhanced. In these simulations, only one cue dimension is selected for enhancement per utterance, although one could imagine a similar scheme in which multiple cues were enhanced, possibly weighted by their current reliability. Here, cues are selected for enhancement based on their distributionally-defined reliability ω_d:

$$\omega_d = \frac{d'(d)}{\sum_{j=1}^{D} d'(d_j)} \tag{3.6}$$

where $d'(d)$ is the difference in means divided by the average standard deviation for the conditional distributions $p(x_d|c_1), p(x_d|c_2)$:

$$d'(d) = \frac{(\mu_{d|c_1} - \mu_{d|c_2})^2}{(\sigma_{d|c_1} + \sigma_{d|c_2})/2} \tag{3.7}$$

Once a specific cue has been targeted for enhancement, its production target value x_d is

modified according to its reliability ω_d, the range of possible values given the category from which it was drawn, and the current global enhancement probability ϵ^β. Enhancement is implemented by sampling from a modified distribution with an exaggerated mean ($\tilde{\mu}$) and a reduced variance ($\tilde{\sigma}$), i.e. a potentially more reliable category exemplar. A multiplier γ_d is calculated as

$$\gamma_d = \exp(\epsilon^\beta)\omega_d \tag{3.8}$$

and this multiplier is used to generate new estimates for $\tilde{\mu}_d$ and $\tilde{\sigma}_d$. This is done to reflect the hypothesis that more informative cues (larger ω_d) are more likely to be produced with extreme values and with less overall variance. The enhanced estimate $\tilde{\mu}$ is based on the distance from the current mean μ_d to one standard deviation above or below the mean, as appropriate, weighted by γ_d.

$$\tilde{\mu}_d = \mu_d \pm \gamma_d |\mu_d - (\mu_d + \sigma_d)| \tag{3.9}$$

The enhanced estimate of the variance $\tilde{\sigma}_d$ is reduced by $\sigma_d \gamma_d$:

$$\tilde{\sigma}_d = \sigma_d - (\sigma_d \gamma_d) \tag{3.10}$$

The ENHANCED PRODUCTION TARGET x'_d is then generated by sampling from a Gaussian with these new parameters $\tilde{\mu}_d, \tilde{\sigma}_d$:

$$x'_d \sim \mathcal{N}(d|k; \tilde{\mu}_d, \tilde{\sigma}_d) \tag{3.11}$$

As a result, more reliable cues are more likely to be produced with extreme values than less realiable cues, and cues will be enhanced to a greater extent when error (ϵ) is high and β is low (i.e., functional load is high).

Bias factors

Finally, the production target may also be modified along one or more cue dimensions by adding or subtracting an appropriate bias term λ_d, used to represent the sum total of channel noise introduced by all phonetic bias factors relevant for a given phonetic dimension (Garrett and Johnson, to appear). In the simulations reported here, the bias term is scaled relative to the distance between category means, reaching 0 when the means become identical.

$$\lambda_d = \log(|(\mu_d|c_2) - (\mu_d|c_1)| + 1) \tag{3.12}$$

Perception

Once the production target $\mathbf{x}$ has been appropriately modified, it is presented to the listener agent for classification. The listener assigns $\mathbf{x}$ a category label c_k with probability $M(c_k|\mathbf{x})$ (Nearey and Hogan, 1986; Ashby and Alfonso-Reese, 1995):

$$M(c_k|\mathbf{x}) = \log \frac{p(x_1|c_k)p(x_2|c_k), \ldots, p(x_D|c_k)p(c_k)}{\sum_{k=1}^{K} p(x_1|c_k)p(x_2|c_k), \ldots, p(x_D|c_k)p(c_k)} \tag{3.13}$$

Once labeled, $\mathbf{x}$ is added by the listener to the appropriate exemplar list, based on the label assigned. (While it is not implemented in this version of the architecture, tokens uttered by the speaker agent could also be added to the speaker's own exemplar list to simulate the influence of self-production on phonetic realization.) Both agents then recompute the memory weights α for each exemplar in the lists $\mathcal{E}_k$ for all K components using Equation (3.1); exemplars with a value of α less than some threshold not used when estimating θ at the beginning of the next iteration.

3.3 Summary

This chapter has described the implementation of an agent-based architecture the may be used for the simulation of sound change. The production and perception of phonetic categories is accomplished using the GMMs introduced in Chapter 2, but the framework also provides an implementation of the PROBABILISTIC ENHANCEMENT HYPOTHESIS – the idea that subphonemic cues are selected and enhanced in accordance with, and in proportion to, their contribution to the successful identification of a phonetic category. The framework also includes parameters to tune the FUNCTIONAL LOAD of a contrast and to introduce BIAS FACTORS on a cue-by-cue basis.

CHAPTER 4

TRANSPHONOLOGIZATION IN SEOUL KOREAN

In this chapter, I will employ the computational modeling framework described in Chapters 2 and 3 to show how a particular case of sound change, the transphonologization of F0 in Seoul Korean, may be understood as the interaction of ADAPTIVE ENHANCEMENT with bias-driven loss of contrast precision, but that depending on the particulars of the situation, the interaction of these factors may not always lead to a reorganization of subphonemic cues. In particular, I will show how the SELECTION and TRADING PROBLEMS introduced in Chapter 1 can be understood by viewing phonologization as the result of bias-driven enhancement, and that on their own, neither the notions of bias (noise) or enhancement (optimization) on their own appear to be sufficient.

4.1 Selection and trading in transphonlogization

Chapter 1 identified three questions raised by the phonologization model: the problems of SELECTION, TRADING, and RESTRUCTURING. Here, we focus on the problems of SELECTION and TRADING (the RESTRUCTURING PROBLEM will be treated in detail in Chapter 5). Solving the SELECTION problem involves determining which cue is likely to be targeted by a phonologization process, while solving the TRADING problem involves explaining why phonologization is so often (perhaps invariably) accompanied by *de*phonologization – that is, why a single cue tends to dominate in the perception of a given phonetic contrast.

The transphonologization of fundamental frequency (F0) in Seoul Korean illustrates both of these problems quite clearly. Recall that while VOT was the primary acoustic-phonetic dimension distinguishing lenis /p t k/ from aspirated /p^h t^h k^h/ voiceless stops as spoken by Seoul Korean adults during the 1960s, F0 has taken over as the primary dimension as

indicated by recordings of Seoul Korean adults taken during the 2000s[1]. This difference is can be seen visually in Figure 4.1, which plots 500 exemplars each of /pa/ and /p^{h}a/ drawn from five normally distributed cues with parameters estimated from published data.

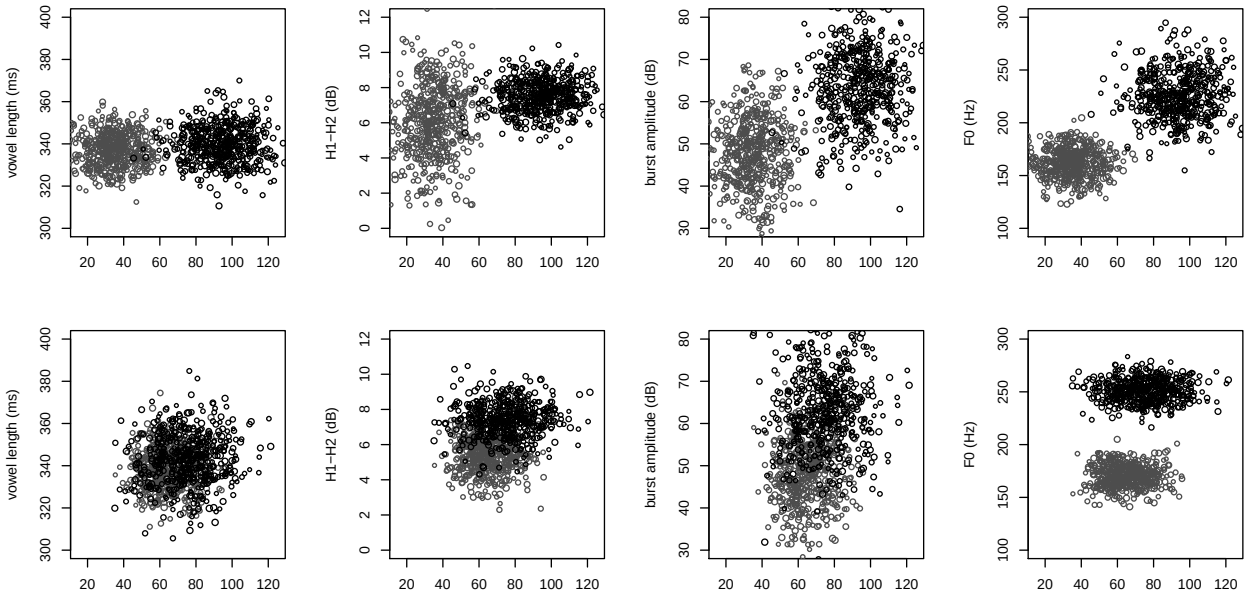

Figure 4.1: Top row: distribution of lenis /p/ and aspirated /p^{h}/ stops, Seoul Korean, 1960s. Bottom row: lenis /p/ and aspirated /p^{h}/ stops, Seoul Korean, 2000s. X axes represent VOT (in ms), y axes represent (left to right) following vowel length (in ms), $H_1 - H_2$ (in dB), burst amplitude (in dB), F0 at vowel onset (in Hz). Based on data from Cho et al. (2002); Kim et al. (2002); Silva (2006a); Kang and Guion (2008).

While it appears from Figure 4.1 that F0 may have already functioned as a redundant cue for speakers born during the 1940s and 1950s, it was not the only redundant cue: even visually, a distinction between the two categories may also be surmised on the basis of $H_1 - H_2$ (column 2) or, to a somewhat lesser degree, closure duration (column 3). So why was F0 selected as the primary dimension along which to maintain the contrast between lenis and aspirated stops, and why did VOT become so uninformative in this regard? Can we better understand the dynamics of how and why such a shift might have taken place?

1. As the production of cues to the fortis series has not changed significantly from generation to generation (Kang and Guion, 2008), only the lenis and aspirated series are illustrated for clarity.

By integrating the mixture modeling framework from Chapter 2 in an agent-based iterated learning environment described in Chapter 3, we can model how the interaction of bias factors and an adaptive enhancement strategy results in changes to individuals' internal representations over time. The Korean results suggest that while phonologization may result from the interaction of probabilistic enhancement with bias factors (i.e., channel noise), neither enhancement nor bias alone is sufficient to explain either the SELECTION or TRADING problems.

This chapter is organized in the following fashion. §4.2 reviews previous work on the phonetics of the Korean laryngeal contrast and the phonologization of F0, and motivates a potential bias factor which may have contributed to a loss of contrast position among the initial obstruents. §4.3 describes the results of a number of computational simulations employing the architecture described in Chapter 3, and §4.4 discusses these results in terms of a functional-adaptive view of sound change.

4.2 The laryngeal contrast in Seoul Korean

This section provides background to the word-initial stop contrast in Seoul Korean and reviews previous work relevant to the phonologization of F0 in this language.

4.2.1 Phonetic cues to the laryngeal contrast in Seoul Korean

Korean is typologically unusual among languages of the world in that it distinguishes between three types of (phonetically) voiceless obstruents: *fortis*[2] (or tense) /p* t* k*/, *lenis* (or lax) /p t k/, and *aspirated* /p^h t^h k^h/ (Sohn, 1999). Based on production data such as that shown in Table 4.1, some authors concluded that VOT was the primary acoustic

2. Following recent work such as Kim et al. (2002); Wright (2007); Kang and Guion (2008), tense stops are represented with a following asterisk as /p* t* k*/, in part because these diacritics do not have any other conventionalized phonetic meaning; however, they have also been represented as /p' t' k'/, /pp tt kk/ and /P T K/. Note that the IPA provides no standard means of distinguishing fortis and lenis stops, although the Extensions do have diacritics for 'strong' and 'weak' articulations.

correlate distinguishing the three types of stops. Lisker and Abramson (1964) found that while there was some degree of overlap between VOT values for categories at the same place of articulation, separation was more pronounced in strong prosodic positions (e.g. sentence-initially compared to sentence-medially). Similar results (for speakers born in the 1940s and 1950s) were replicated in studies by C.-W. Kim (1965) and Han and Weizman (1970).

isolated words									
	p*	p	p^h	t*	t	t^h	k*	k	k^h
$\bar{x}$	7	18	91	11	25	94	19	46	126
range	0–15	10–35	65–115	0–25	15–40	75–105	0–35	30–65	85–200
n	15	30	21	16	24	12	16	34	12
sentence-initial									
	p*	p	p^h	t*	t	t^h	k*	k	k^h
$\bar{x}$	7	22	89	11	30	100	20	48	125
range	0–15	15–40	55–115	0–20	15–40	75–130	0–30	35–75	80–175
n	14	28	24	15	21	12	14	35	10
sentence-medial									
	p*	p	p^h	t*	t	t^h	k*	k	k^h
$\bar{x}$	5	13	75	12	22	78	21	44	93
range	0–10	10–20	40–130	0–25	10–45	50–120	10–35	30–65	55–175
n	14	10	23	16	12	10	14	11	10

Table 4.1: Korean VOT data from Lisker and Abramson (1964), from a single Seoul Korean speaker of unknown age and gender. Durations are listed in milliseconds (ms).

There is also articulatory evidence pointing to differences between the various types of Korean voiceless stops. Lee and Smith (1972) showed that subglottal pressure for the aspirated series was higher than for the other two stop types. Using an electromyograph (EMG) in conjunction with fiberoptic observation, Hirose, Lee, and Ushijima (1974) investigated the intrinsic actions of the laryngeal muscles during the production of Korean stops and found that the period of glottal width during stop closure to be narrower for fortis and lenis stops

compared to aspirated stops, accompanied by a slight increase in activity of the interarytenoid muscle. They also found patterns of vocalis and lateral cricoarytenoid muscle activity characteristic of the fortis series, presumably the result of 'an increase in inner tension of the vocal folds as well as...constriction of the glottis during or immediately after the articulatory closure' (1974:151), while aspirated stops were characterized by a marked suppression of all adductor muscles immediately prior to release. These results suggest that that the production of Korean stops involves differences of glottal tension as well as glottal width.

The Korean initial stops also differ in terms of the intensity of the release burst (Kim, 1965; Han, 1996; Cho et al., 2002). For instance, Cho et al. (2002) found the relative energy (amplitude) associated of the stop burst to vary with stop type in Korean, with aspirated stops having considerably higher energy than fortis or lenis stops. The authors attribute these differences to the aerodynamic effect of intraoral airflow rate: while both the aspirated and fortis Korean stops have comparable intraoral air pressure, fortis stops are produced with reduced intraoral airflow, possible a result of increased glottal resistance caused by greater tension in the vocal tract during production of this stop. While lenis stops have greater intraoral airflow than fortis stops (but less than aspirated stops), they are also produced with relatively low intraoral air pressure.

Differences in fundamental frequency (F0) have long been noted for Korean stops. Han and Weizman (1970) reported that the average F0 for aspirated and fortis stops was higher than that of lenis stops (Table 4.2). Similar results were reported by Kagaya (1974), Hirose, Park, Hirohide Yoshioka, and Umeda (1981), and Cho et al. (2002). These differences are considerably exaggerated when compared to the well-known intrinsic effect of F0 perturbation induced by stop type. Kim (2000) compared the F0 perturbations following Korean lenis vs. fortis/aspirated stops with those following English voiced vs. voiceless stops and found that while the F0 differences between vowels following English stops were negligible by vowel midpoint, F0 differences induced by the Korean stops extended over the course of

the entire vowel, with F0 of vowels following fortis and aspirated stops 30 Hz higher than F0 of vowels following lenis stops at vowel offset (see also Jun, 1993, 1996; Ahn, 1999).

		p*	p	p^h
speaker 1	$\bar{x}$	178	144	185
(male)	*range*	145–215	130–170	150–220
speaker 3	$\bar{x}$	308	266	341
(female)	*range*	278–322	250–292	318–369

Table 4.2: F0 at vowel onset from two Korean speakers. From Han and Weizman (1970).

The articulatory mechanism which gives rise to this correlation was the subject of considerable research activity, for while the high F0 for aspirated stops has a clear phonetic precursor – in that a high rate of airflow gives rise to an increased Bernoulli effect leading to an increase in the rate of vocal fold vibration (Ladefoged, 1973) – the heightened F0 after tense stops was rather more puzzling: given that the state of the glottis in the production of tense stops in Korean bears many of the hallmarks of creaky voice (Laver, 1980), tense stops might be expected to have a lower F0, even lower than that of lenis stops. However, as noted by Ahn (1999), the high F0 associated with tense stops finds an explanation in the heightened activations of the thyroarytenoid and cricothyroid muscles found by Hirose et al. (1981): the joint action of these muscles increases vocal fold stiffness, promoting high F0, as does the movement of the thyroid cartilage induced by contraction of the cricothyroid. Since both of these gestures are independent of the adduction of the arytenoid cartilages, a characteristically creaky glottal state can occur simultaneously with high F0. This is similar to the theory advanced by Kingston (2005) advances a similar theory to explain the fact that the historical loss of laryngealized (creaky voiced) segments in Athabaskan languages resulted in both high and low tones on the preceding vowel in different daughter languages.

Another vocalic effect that has been reported for Korean is that duration of the vowel following the stop is inversely related to degree of aspiration. For instance, Cho (1996)

reports that vowels following tense tops are longer than those following lax stops, which in turn are longer than those following aspirated stops. Table 4.3 gives the the means and ANOVA results for Cho's three male speakers, born in the 1950s or early 1960s, in two conditions: normal and 'slow' (careful) speech. Comparable results for a female speaker born in the late 1960s or early 1970s are reported by Kim et al., shown in Table 4.4, which gives both the duration of the vowel /a/ as well as the duration of the entire CVN syllable in which it occurred. These authors provide similar data for the vowel /i/ and alveolar and velar stops, noting that while the results are broadly in agreement with Cho's data, the lenis and aspirated categories overlap considerably (2002:82).

	Informant 1		Informant 2		Informant 3	
	normal	*slow*	*normal*	*slow*	*normal*	*slow*
p*	155.39	420.18	126.65	449.99	179.76	477.41
p	152.05	401.05	107.02	451.14	131.02	439.73
p^h	124.73	393.68	80.20	410.88	89.41	407.77
p	.0003	.1796	.0001	.0045	.0084	.0001
F	11.784	1.8459	17.907	6.8193	5.8665	16.143

Table 4.3: Mean vowel length (in ms) following fortis, lenis, and aspirated bilabial Korean stops in two conditions. After Cho (1996).

	p*	p	p^h
vowel duration	168(166-171)	157(143–177)	110(100-117)
syllable duration	337(322-355)	369(337–388)	339(327–356)

Table 4.4: Vowel and total syllable duration (in ms) of the vowel /a/ following fortis, lenis, and aspirated stops, in the format *mean(range)*. From Kim et al. (2002).

Finally, phonation type (manifested as relative harmonic intensity, or $H_1 - H_2$) has been argued by several authors to be a reliable indicator of stop type in Korean, or at least to distinguish the tense stops from other types of stops (Ahn, 1999; Kim et al., 2002; Kim and

Duanmu, 2004). Fiberoptic studies such as Kim (1965) and Kagaya (1974) observed that glottal apertures at voice onset were comparable for the lenis and aspirated stops, whereas fortis stops evidenced complete contact at voice onset. The expected acoustic consequence of this difference would be a more dominant first harmonic (H_1) for lenis and aspirated stops as compared to tense stops (Stevens, 2000). In terms of a normalized $H_1{}^*-H_2{}^*$ measure, Ahn (1999) found significant differences between fortis and both lenis and aspirated stops, but not between lenis and aspirated stops. Conversely, Cho et al. (2002) found that $H_1 - H_2$ values for fortis, lenis, and aspirated stops all differ significantly from one another in *post hoc* pairwise comparisons. The results of the Kim et al. (2002) study, given in Table 4.5, provide further support for an $H_1 - H_2$ difference between all three stop types.

	fortis	lenis	aspirated
labial	$-6.2(-6.6$ to $-6.0)$	$1.2(-0.9$ to $3.5)$	4.3(2.1 to 6.0)
velar	$-7.1(-9.1$ to $-6.0)$	4.0(3.5 to 4.8)	$0.2(-4.3$ to $3.3)$

Table 4.5: Mean difference (in dB) in the amplitude of the first and second harmonics ($H_1 - H_2$) at vowel onset following fortis, lenis, and aspirated stops at two places of articulation, in the format *mean(range)*, for a single female speaker of Seoul Korean. From Kim et al. (2002).

Based on the studies reviewed above, it is clear that multiple temporal and spectral properties serve to differentiate fortis, lenis, and aspirated stop in Seoul Korean. Although some of these properties are events that occur during the production of the stop itself and some occur during the production of the following vowel, we might reasonably consider them all to be cues to the stop contrast (and not the vocalic contrast) as they persist to varying degrees regardless of the quality of the following vowel. More convincing still, however, are demonstrations of the perceptual relevance of these significant differences in cue production.

4.2.2 Perceptual studies of the Korean laryngeal contrast

Kim (1965) argued that VOT differences alone could not be the sole characteristic distinguishing fortis from lenis stops in Korean, due to the still considerable degree of overlap even in strong prosodic positions. This was supported by subsequent perceptual studies: Han and Weizman (1970) found that manipulating VOT using edited natural /p^h/ stimuli cued only a two-way lenis/aspirated distinction; Abramson and Lisker (1972), employing a synthetic VOT continuum, found inconsistent response patterns across listeners (while some reported a three-way distinction, others reported only two, while one unexpectedly reported tense stops in the middle of the continuum).

The failure of VOT to unilaterally distinguish between stops led researchers to investigate the potential role of other acoustic dimensions in the perception of Korean stops. Han (1996) created a continuum of F0 at vowel onset over a range of 35 Hz, and found that as F0 increased, listeners were more likely to identify the preceding stop as fortis (vs. lenis). Francis and Nusbaum (2002) investigated how native Korean speakers weight vocalic and consonantal cues to the initial stop contrast as part of a study on the acquisition of novel (L2) phonetic contrasts. They determined that native speakers make phonetic decisions about natural stimuli based not only on VOT, but also on F0 at vowel onset, the clarity of formant structure at the onset of phonation, and the elapsed duration from the onset of vocalic formant activity to the peak vowel amplitude.

Cho (1996) asked listeners to identify initial stops on the basis of CV syllable fragments with a constant vocalic portion from which all aperiodic information and the first two pitch periods had been removed (to avoid including any coarticulatory information in the form of e.g. formant transitions). Overall identification accuracy across stop types was 67%, but accuracy for aspirated stops was at chance, suggesting that vocalic portion of the signal holds particularly strong cues to the distinction between fortis and aspirated stops. This paradigm was replicated and extended by Kim et al. (2002) in two experiments. In the

first, stimuli were created by cross-splicing the vocalic portion of one syllable (e.g., the /a/ from fortis-onset /p*a/) with the consonantal portion of another (e.g., the aspirated /p^{h}/ from /p^{h}a/). When presented with conflicting cues as to the identity of the initial stop, participants tended to base their decision more on vocalic information such as $H_1 - H_2$ and F0 than on cues located in the consonantal portion: a consonant followed by a 'lenis vowel' was generally perceived as lenis, regardless of its other spectral or temporal properties, for instance. In the second experiment, Kim et al. investigated whether information in the vocalic portion of the syllable *alone* could serve as a cue to the identity (or at least phonation type) of the initial consonant by present listeners with only the relevant vocalic portions of the signal. While the presence of a 'lenis vowel' was generally sufficient to recover a deleted lenis stop, in the absence of consonantal cues such as VOT, fortis and aspirated stops were often confused and labeled as fortis.

4.2.3 Changes in the production and perception of Korean stops

The perceptual studies reviewed above suggest that Korean listeners attend to multiple cues when making decisions about the identity of an initial stop, and that neither consonantal nor vocalic information alone is sufficient to convey the three-way distinction between initial stops. However, as Kim et al. note, the distinction between lenis stops and 'other' stops is quite robustly cued by F0 at vowel onset, which is 'sufficiently perceptually robust as to uniquely specify lax stops in the absence of any consonantal information prior to voice onset' (2002:99). A comparison of studies conducted during the 1960s with those conducted 30 to 40 years later suggests that this is a more recent development, accompanied by a decrease in the informativeness of VOT as a cue distinguishing lenis from aspirated stops. Speakers born in the 1960s tend to produce lenis stop with more aspiration (∼40-70ms) while aspirated stops are produced with slightly less (∼85-105ms) compared to speakers born in the 1940s and 1950s (Kim, 1965; Han and Weizman, 1970; Silva, 1993; Cho et al., 2002; Wright, 2007).

This trend has now been well-established by a number of apparent time production studies including those of Silva (2006a), Wright (2007), and Kang and Guion (2008). Silva (2006a) reports VOT and F0 data from recordings of 36 adult native Seoul Korean speakers born between 1943 and 1982. He found that the degree to which VOT serves to differentiate lenis from aspirated stops decreases as a quadratic function of age. Wright (2007) examined VOT and closure duration of aspirated and lenis stops in the speech of 20 Seoul Korean speakers born between 1955 and 1987. In the aggregate, older speakers tended to produce lenis and aspirated stops with distinct VOT and closure duration values, whereas these cues were produced with more similar or even identical values by younger speakers.

However, as VOT has become increasingly less informative in distinguishing between lenis and aspirated stops in Seoul Korean, F0 appears to have become more informative. Silva (1992) demonstrated that differences in F0 appeared to be being exaggerated as differences in VOT diminished – a clear instance of theTRADING PROBLEM. While Silva (2006a) failed to find an effect of speaker age on F0 values (although F0 did generally distinguish lenis from fortis and aspirated stops for speakers of all ages), the apparent time study by Kang and Guion (2008) presents evidence that younger and older Korean speakers do in fact treat cues to the initial stop contrast differently. In their first experiment, Kang and Guion sought to establish that younger and older Seoul Korean speakers produce cues to the initial laryngeal contrast differently, and to ascertain whether these differences are more pronounced in some speaking styles than in others. The results of this experiment are shown in Figure 4.2. Older speakers made clear distinctions between lenis and aspirated stops in terms of VOT in conversational, citation, and clear speech conditions, whereas younger speakers only made this distinction in the clear speech condition, and then only to a very small degree. Kang and Guion also examined differences in the production of $H_1 - H_2$ and F0 between age groups and across speech conditions. While $H_1 - H_2$ productions varied somewhat by speech condition, with differences between fortis and other stops begin somewhat more exaggerated

in clear speech, the differences were largely comparable across age groups. For F0, on the other hand, an interaction between speech condition and age group was observed: younger speakers tended to expand the F0 distance between lenis stops and fortis/aspirated stop types, especially in clear speech.

In a second experiment, Kang and Guion tested whether the expanded F0 productions observed for the younger group were for the purpose of enhancing the phonological lenis vs. aspirated contrast, or simply the result of an expanded F0 range in clear speech. They found that while the difference between the stops /t/ and /t^h/ increased substantially in clear speech for younger speakers, the difference in F0 between the segments /h/ (a 'high-tone trigger') and /n/ (a 'low-tone trigger': Jun, 1993) was not particularly greater in clear speech as compared to citation-form speech. This suggests that the expanded F0 range between lenis and aspirated stops for younger speakers is related to maintenance of the phonological contrast between lenis and fortis/aspirated stops, rather than simply as e.g. an aerodynamic by-product of some other articulation.

In short, the empirical facts of subphonemic sound change affecting lenis and aspirated stops in clear speech conditions in Seoul Korean most relevant to the present work can be summarized as follows:

1. Lenis stops are produced with greater VOT, and aspirated stops are produced with reduced VOT, compared to productions in similar speech conditions by older speakers.

2. The difference in F0 between lenis stops and fortis/aspirated stops is increased in clear speech in just those cases where it helps to differentiate the phonological stop contrast.

One account of this shift is that as VOT became less informative in terms of helping to categorize these stops, F0 became more important. However, this raises at least two further questions: first, why VOT summarily became less informative in the first place (the TRADING PROBLEM) as well as why F0, and not some other cue (say, closure duration, or $H_1 - H_2$), became the primary indicator of the phonological contrast (the SELECTION PROBLEM).

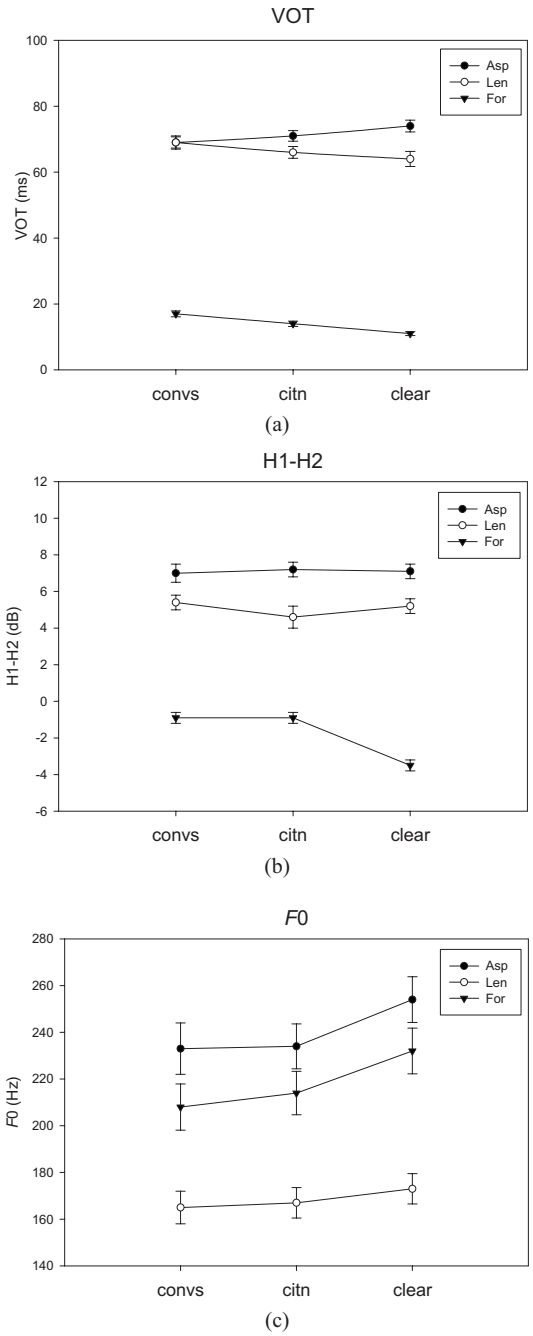

FIG. 1. Mean values with standard errors for the production of Korean stops (aspirated, lenis, and fortis) in conversation, citation-form, and clear speech styles by the younger group (n=11) for three acoustic correlates [(a) VOT, (b) H1-H2, and (c) $F0$].

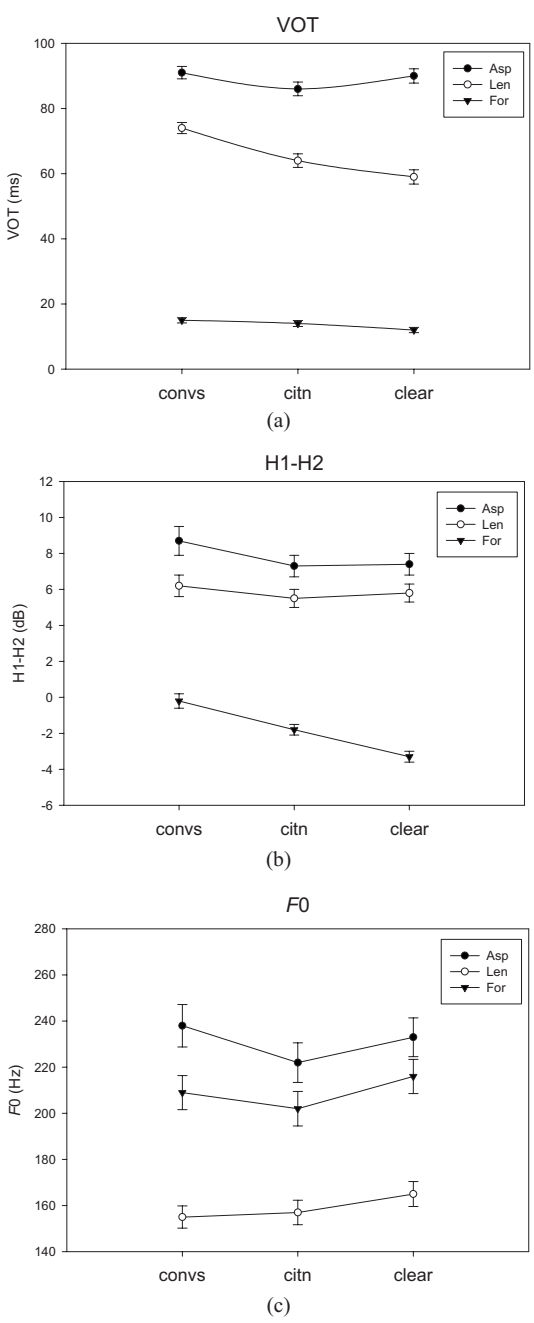

FIG. 2. Mean values with standard errors for the production of Korean stops (aspirated, lenis, and fortis) in conversation, citation-form, and clear speech styles by the older group (n=11) for three acoustic correlates [(a) VOT, (b) H1-H2, and (c) $F0$].

Figure 4.2: Figures 1 and 2 from Kang and Guion (2008) showing the differences in the production of VOT, $H_1 - H_2$, and F0 in three speech conditions for a group of younger speakers (Fig. 1, column 1) compared to a group of older speakers (Fig. 2, column 2).

Although historically shown to employ a pitch-based accentual system, most varieties of modern Korean are not considered tonal in the usual sense (see Silva, 2006a and references therein). However, other dialects, such as the Kyungsang dialects spoken outside of Seoul, have been argued to retain the Middle Korean tone system (Chung, 1991; Chang, 2007; Hong, 2007). Given the influx of immigrants to Seoul in the years following the Korean War, contact with a tonal dialect may have precipitated the shift in the production of cues among Seoul speakers. I will not explore that hypothesis here, but the architecture developed in the remainder of this chapter could certainly be used to simulate such a scenario.[3]

4.2.4 Phonetic bias factors in the production of Korean stops

In terms of accounting for the increasingly similar VOTs of lenis and aspirated stops produced by younger Seoul Korean speakers, one account in particular points to systemic phonological factors as a possible source of phonetic production bias. Silva (1992, 1993, 2006a) makes the case that 'any non-sonorant singleton consonant...in phrase initial position is realised with aspiration' (2006a:302) by demonstrating that the duration of VOT during the production of 'non-sonorant singleton' stops (i.e. non-fortis) at the edge of a prosodic phrase is considerably longer (∼60ms) than at the edge of a word (∼20ms) or word-internally (∼3ms: see Table 4.6). This may have led to lenis and aspirated stops being produced with similar VOT in initial position, which would lead to increased distributional overlap over time. On Silva's account, fortis stops would not be subject to this same bias, since they are phonologically geminate (2006a:303).[4]

That Korean fortis stops may be treated as phonologically geminate, or in any event dif-

3. Related to this is the question of who to determine whether or not a language has a lexical tone contrast or something else, or when a phoneme-level prosodic distinction has become lexicalized. I will not pursue this question here, but see Wright (2007); Hyman (2009) for some recent discussion and references.

4. Since this proposed bias factor would not have affected the production of fortis stop series, the following simulations focus on the lenis/aspirated contrast in the interest of expository clarity.

	ϕ-Edge	ω-Edge	ω-Internal
Voicing during closure	10 ms	17 ms	33 ms
% of closure that is voiced	23%	36%	77%
Post-release VOT	60 ms	22 ms	3 ms

Table 4.6: Degree of voicing during closure and post-closure release aspiration (VOT) of Korean lenis stops in three prosodic positions: minor-phrase (ϕ) edge, word (ω) edge, and word-internal. From Silva (1993).

ferently from non-fortis stops, is also supported by acoustic data. In examining the recordings of two subjects in conjunction with electromyographic data, Hirose et al. (1981) found that the closure period of word-initial fortis stops were considerably longer than those of lenis or aspirated stops (Table 4.7). The function of this length increase may be to build up air pressure behind the closure, in order to compensation for the relatively small glottal opening characteristic of Korean fortis stops.

		k*	k	k^h
speaker P	$\bar{x}$	207	146	145
	s.d	1.1	1.1	0.7
	n	15	15	16
speaker C	$\bar{x}$	150	115	143
	s.d	0.5	0.7	0.7
	n	12	12	12

Table 4.7: Duration of stop closure (in ms) for word-initial velar stops /k* k k^h/ from two Seoul Korean speakers (n = number of tokens). Adapted from Hirose et al. (1981).

There is also some evidence of a gender imbalance in the production of VOT. Data from 6 Seoul Korean speakers born in the 1960s and 1970s collected by Kim (1994) suggests that female speakers may have begun to produce lenis and aspirated stops with similar VOT patterns prior to males (Table 4.8). While certainly not conclusive, these data suggest a

possible pathway through which this innovation may have spread to a younger generation (see e.g. Labov, 1990, 2001 for more on the role of women as leaders of linguistic change).

	p*	p	p^h
male	8.7(3.2)	45.6(18.2)	76.7(18.2)
female	7.8(1.9)	77.7(18.4)	71.2(20.7)

Table 4.8: Means and standard deviations (in ms) of VOT data based on 3 male and 3 female speakers of Seoul Korean, aged 25–35 (born 1962–1974) at the time of data collection. Adapted from M.-R. Kim (1994).

Together, these studies suggest higher-level factors which may be exerting a bias on the phonetic realization of VOT in Seoul Korean in the production of lenis and aspirated stops. This bias factor will be crucial in explaining the shift in relative reliability from VOT to F0 in this language.

4.2.5 An adaptive account of sound change in Seoul Korean

From the studies reviewed above, it is clear that both the production and perception of cues to the Korean stop contrast have changed over time, possibly precipitated by the influence of a particular phonetic bias factor. Without loss of generality, we might arbitrarily divide the period from the 1960s to the 2000s into two periods, a 'before' and an 'after'. Our task then becomes to explain the role played by this the bias factor in the subphonemic evolution of the contrast.

As reviewed in Chapter 1, adaptive theories such the H & H theory (Lindblom, 1990; Lindblom et al., 1995) hypothesize that speakers modify their productions in response to their estimates of the listener's perceptual needs. If the listener's need is estimated to be high, the speaker is more likely to enhance (hyperarticulate) some aspect of the speech signal. The prosodically-conditioned VOT bias discussed above may have conditioned a loss in the precision of the lenis/aspirated stop contrast in Korean, and the cue targeted for

enhancement clearly seems to have shifted from VOT to F0 over time. But why F0, as opposed to some other cue? Cue selection can be (at least probabilistically) predicted in a model in which the representation of speech sound categories allows for reliability of cues to be quantified (such as the GMM representation laid out in Chapter 2). Based on their assessment of the listener's needs, an adaptive speaker will, by hypothesis, enhance cues in proportion to their reliability: the more reliable a cue dimension, the more likely it will be enhanced.

In the Seoul Korean case, that cue would appear to be VOT. Table 4.9 compares the cue reliability ω for five of the cues discussed in §4.2.1 above – note that F0 ($\omega = 0.32$) is (was) nearly as reliable as VOT ($\omega = 0.4$). This state of affairs suggests at least three hypotheses as to how the sound change might have occurred. First, if cue enhancement is probabilistic, rather than deterministic, it is conceivable that F0 may have been enhanced often enough, or by e.g. socially prominent enough speakers, to become more reliable than VOT over time. On this account, the sound change would be driven entirely by (probabilistic) enhancement. Alternatively, if the reliability of VOT were somehow reduced – as a result of the phonetic bias discussed above, for instance – the relatively reliability of F0 would increase (as ω is normalized measure). Finally, the transphonologization of F0 may have been the result of some interaction of these two factors.

	Category	VOT	VLEN	H_1-H_2	BA	F_0
1960s	lenis	35 (11)	337 (8)	6 (2)	48 (8)	162 (14)
	aspirated	93 (15)	340 (15)	7.5 (1)	64 (9)	227 (21)
	ω	0.4	0.03	0.09	0.16	0.32

Table 4.9: Parameter values and weights for cues to Korean stops among the older (1960s) generation, taken or estimated from data in Cho (1996), Kim et al. (2002), Silva (2006a), and Kang and Guion (2008). Standard deviations are given in parenthesis. VOT = voice onset time; VLEN = vowel length; BA = burst amplitude; F_0 = F0 at vowel onset.

4.3 Simulating phonologization in Seoul Korean

Data from the apparent time studies reviewed above suggest that while the distinction between lenis and aspirated stops in Seoul Korean of the 1960s was mainly cued by VOT, this distinction is now cued chiefly by F0 of the following vowel and has been accompanied by a loss of contrast along the VOT dimension. As detailed in §4.2.5, this kind of change is consistent with an adaptive theory of phonetic enhancement, whereby a shift in cue weights involves loss of contrast along a once primary cue dimension concomitant with (probabilistic) enhancement along a previously redundant dimension. In what follows, it will be shown that probabilistic enhancement accurately predicts both the cue enhanced (F0) as well as the degree of enhancement in Seoul Korean.

In order to explore how language users choose among competing cues in phonologization (the SELECTION PROBLEM), the Korean stop contrast was modeled in five acoustic-phonetic dimensions: voice onset time (VOT), F0 and duration of the following vowel, the difference in amplitude between the first two formants of the vowel ($H_1 - H_2$), and the amplitude of the release burst. Agents' lexica consisted solely of the syllables /pa/ and /$\mathrm{p^h a}$/[5]. In each simulation, two ideal observer agents were seeded with initial distributions for each of these cues estimated from data reported in Cho (1996), Kim et al. (2002), Silva (2006b), and Kang and Guion (2008) given in Table 4.10; two-dimensional scatterplots showing the joint distributions of VOT (the primary cue) and each of the other cues are given in Figure 4.3. Table 4.10 also gives the distributions and reliability scores ω for these five cues for younger speakers of Seoul Korean, which were estimated from the above sources. Since increased reliance on F0 in Korean has been accompanied by a reduction in reliance on VOT (the TRADING PROBLEM), the second goal of the simulations was to see if it was possible to replicate this transition using the current implementation, and if so, under what conditions.

5. Although no contextual or lexical effects were considered in these simulations, this should not be taken to imply that these effects play no role in the process of sound change; the focus of the present study was simply on the subphonemic dynamics.

In order to simulate the effects of the prosodically-conditioned aspiration bias described in §4.2.4, the bias factor λ implemented here affected the VOT dimension only.

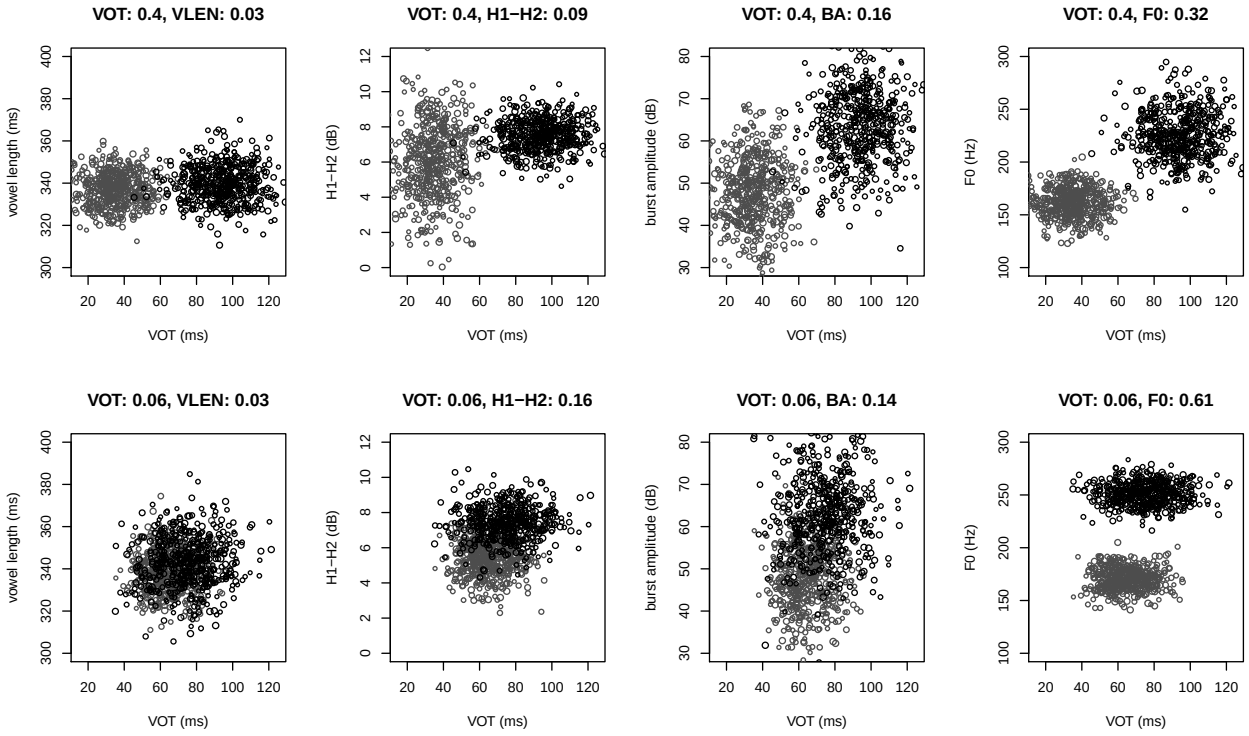

Figure 4.3: Row 1: distribution of five cues to the laryngeal contrast in Korean used to seed the simulations. Row 2: modern distribution of the same cues. Data estimated from Cho (1996), Kim & Beddor (2002), Silva (2006a), Kang and Guion (2008). Captions give cue reliability ω as computed by Equation (2.17). VOT = voice onset time; VLEN = vowel length; BA = burst amplitude.

Three series of simulations are reported, each seeded from the same initial parameterization. The first round of simulations considered the effects of applying a probabilistic enhancement strategy without systemic bias (§4.3.1). The second round of simulations considered the effect of applying systemic bias to the production of a single cue, but without enhancement (§4.3.2). The third round of simulations explored the effects of applying both systemic bias and probabilistic enhancement (§4.3.3). Simulations were run for varying lengths of time; those reported here are representative runs of 25,000 iterations, at which

	Category	VOT	VLEN	H_1-H_2	BA	F_0
1960s	lenis	35 (11)	337 (8)	6 (2)	48 (8)	162 (14)
	aspirated	93 (15)	340 (15)	7.5 (1)	64 (9)	227 (21)
	ω	0.4	0.03	0.09	0.16	0.32
2000s	lenis	65 (11)	338 (10)	5.5 (1)	48 (8)	170 (10)
	aspirated	73 (15)	343 (12)	7.5 (1)	64 (9)	250 (11)
	ω	0.06	0.03	0.16	0.14	0.61

Table 4.10: Parameter values and weights for cues to Korean stops, taken or estimated from data in Cho (1996), Kim et al. (2002), Silva (2006a), and Kang and Guion (2008). Standard deviations are given in parenthesis. VOT = voice onset time; VLEN = vowel length; BA = burst amplitude.

point neutralization along the primary cue dimension tended to become complete and/or probability of enhancement approached zero. All simulations reported here utilized an exemplar pruning threshold of 0.3 and a value for the β exponent of 0.5.

In order to quantify the goodness of fit between the target distributions and the results of the various simulations, the relative entropy or KULLBACK-LEIBLER DIVERGENCE (KL divergence: Kullback and Leibler, 1951) between each target and simulated cue dimension was calculated. Since the cue distributions here are continuous, KL divergence between distributions F and G with densities $f(x)$ and $g(x)$ is estimated as

$$KL(F||G) = \int_{-\infty}^{\infty} f(x) \log \frac{f(x)}{g(x)} \, dx \tag{4.1}$$

KL divergence provides a (non-symmetric) measure in bits of the dissimilarity between two distributions; $KL(F||G)$ is equal to zero when two distributions F and G are identical, and grows with the dissimilarity between them. Although a multivariate formulation of KL divergence also exists, summed univariate distances are reported here in keeping with the overall independence assumptions.

4.3.1 Enhancement, no bias

Based on the cue weights in Table 4.10, it would appear that a contrast along the F0 dimension already existed in Seoul Korean of the 1960s, albeit covertly, in that while it may have played a role in native speaker perception of the contrast, it may not have been obvious even to the trained ear of linguists (Scobbie et al., 2000; Yu, to appear). One interpretation of Hyman's characterization of phonologization is that active enhancement of cues on the part of speakers itself conditions the transition of a cue from covert to overt indicator of contrast: in other words, an existing contrast is neutralized *as a result of* a covert contrast becoming overt. One way to test this assumption is to consider the application of probabilistic enhancement in the absence of any external bias. If enhancement alone drives phonologization, then as one cue dimension becomes more informative, another should become less informative.

The enhancement-without-bias simulations proceeded as follows. At each iteration i, a cue dimension d was enhanced with probability $p(\epsilon^{\beta}\omega_d)$ by sampling d from a distribution with parameters $\tilde{\mu}_d, \tilde{\sigma}_d$ as described in §3.2.2 above. The results of a representative simulation run are shown in the first row of Figure 4.4. In all such simulations, the cue with the highest initial reliability ω (here, VOT) maintained its relative dominance throughout. The degree of enhancement was extremely small, reflecting the fact that the precision of the contrast is never jeopardized, although as shown in Figure 4.5, the error rate is generally seen to increase. These simulation results suggest that probabilistic enhancement alone, at least enhancement driven by loss of contrast precision, is probably insufficient to induce phonologization of an existing covert contrast, although if two cues were more or less equally reliable, an enhancement-only model would predict each of them to become dominant with roughly equal frequency. In addition, the enhancement-only model makes the prediction that enhancement along one cue dimension need not entail a reduction in the reliability of another.

4.3.2 Bias, no enhancement

It is also worth considering the inverse of the scenario described above. If a covert contrast already exists, it is possible that enhancement is unnecessary, and that a redundant, secondary cue will simply become primary through the continuous application of systemic bias to the appropriate cue dimension. To test this hypothesis, simulations were run in which a bias was applied to productions of VOT at each timestep, such that VOT values for category c_2 (e.g., [p$^{\text{h}}$]) were produced with slightly shorter VOTs, while values for category c_1 (e.g., [p]) were produced with slightly higher VOTs. The motivation for this bidirectional bias is as follows. In many instances where VOT productions are biased, only one of the two or more VOT distributions change as a result: in a study of the effects of speech rate on stop production, Kessinger and Blumstein (1997) found that long lag and prevoiced VOTs, but not short lag VOTs, changed as a function of speaking rate for speakers of Thai, French, and English; Kang and Guion (2008) observed similar effects for Korean. However, in the modern Korean case, average VOT values of lenis and aspirated stops have clearly converged at a midpoint; it is not the case that lenis stops are produced as aspirated or vice versa. The bidirectional bias factor implemented here is consistent with the possible precursor of this convergence discussed in §4.2.4. Values were modified as in Equation (3.12). In these simulations, no enhancement was applied.

The results of a representative simulation run are shown in the second row of Figure 4.4. The VOT contrast has been neutralized, and due to the normalization of cue weights, F0 has emerged as the primary indicator of the contrast. However, this is slightly different from the attested modern Korean situation, in part because the F0 distributions have not changed: F0 is the most informative cue simply because all other cues have been made highly uninformative. This suggests that a phonetic bias factor can result in the phonologization of a previously redundant cue, so long as the distribution of cue was already sufficient to distinguish between the relevant categories. In the empirical distributions for modern

Korean, however, the category means for both aspirated and lenis obstruents have shifted slightly away from one another, suggesting that they have been enhanced (Figures 4.2 and 4.3, row 2), and indeed Kang and Guion (2008) found different patterns of enhancement behavior for younger and older Korean speakers with respect to the cues they targeted. Thus, while the covert F0 contrast may have been exposed by a systemic production bias, it seem unlikely that this was the only factor at play in this transphonologization.

4.3.3 Bias and enhancement

The third series of simulations considered the effect of applying both systemic VOT bias and probabilistic enhancement. Representative simulation results are shown in the third row of Figure 4.4. These results most closely resemble the empirical results, as seen in Table 4.11. Note that continued application of systemic bias has again resulted in the VOT dimension becoming completely uninformative. While overall F0 variance has been reduced, it still falls short of the empirical variance along this dimension, suggesting a more aggressive model of variance reduction may be warranted.

Note that while the KL divergences reported in Table 4.11 are generally quite small, so too are the the KL divergences between the initial and final (target) distributions, as seen in the last line of Table 4.11. The KL divergences for various dimensions should thus not be interpreted in an absolute sense, but instead relative to other values for the same dimension.

4.3.4 Summary

Table 4.11 gives an overview of the combined simulation results, showing the distribution parameters and weights for each cue after a representative simulation run. The empirical targets, repeated from Table 4.10, are given at the bottom. Figure 4.4 gives a graphical overview of the results, while Figure 4.5 compares contrast precision (ϵ) at simulation timestep for each of the three scenarios.

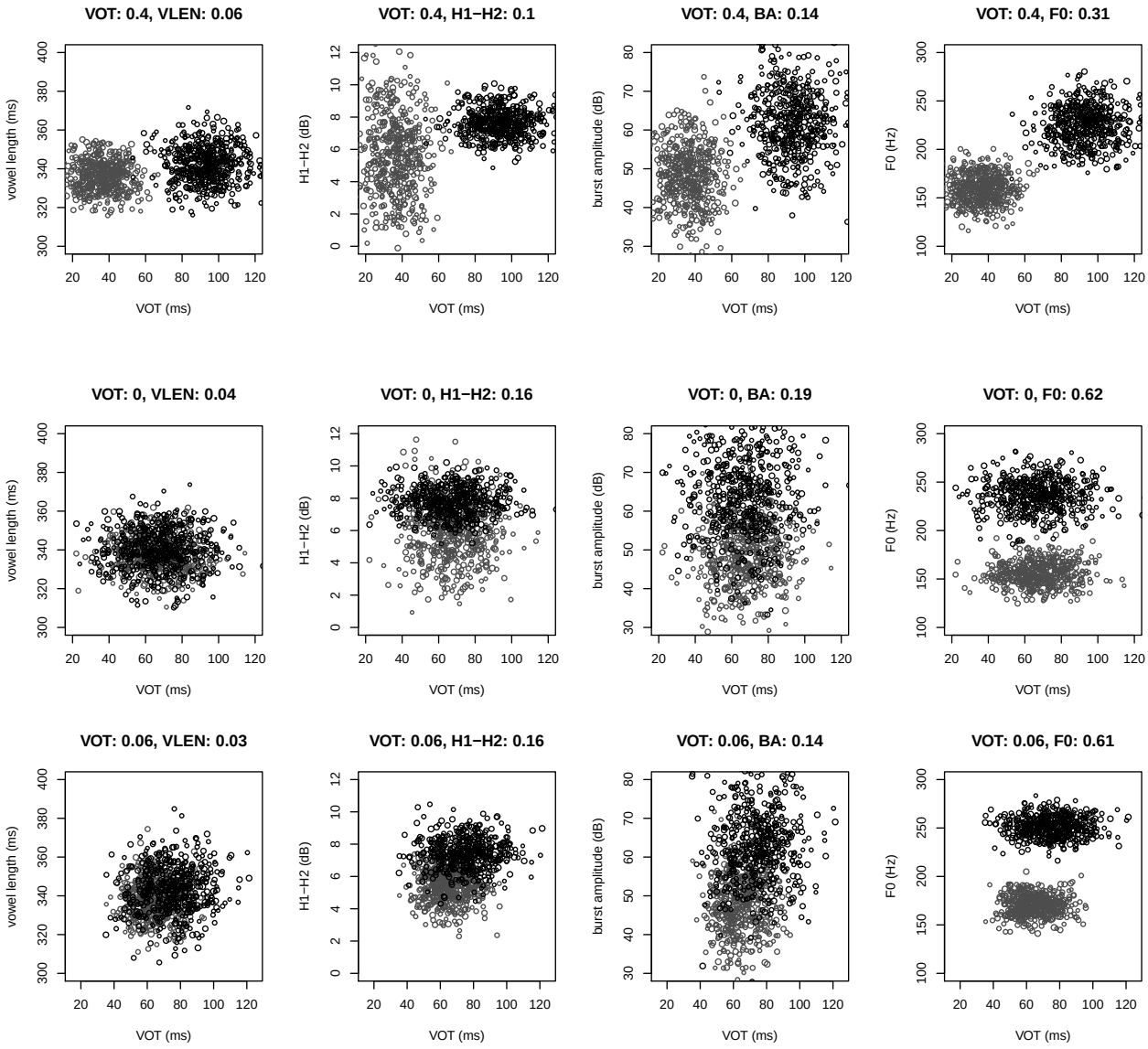

Figure 4.4: Cue distributions after 25,000 iterations for lenis /p/and aspirated /p^h/ stops. Row 1: enhancement without bias. Row 2: bias without enhancement. Row 3: bias and enhancement. Row 4: empirical targets. Captions give cue reliability ω as computed by Equation (2.17).

Source	Category	VOT	VLEN	H_1-H_2	BA	F_0
	lenis	36 (10)	336 (8)	5.6 (2.4)	48 (7.4)	159 (15)
enhancement only	aspirated	92 (13)	342 (10)	7.6 (0.9)	62 (8.7)	225 (20)
	ω	0.4	0.06	0.1	0.14	0.31
	KL	0.2	0.002	0.27	0.05	0.01
	lenis	65 (11)	340 (8)	6.3 (1.8)	48 (7)	162 (12)
bias only	aspirated	65 (16)	340 (9)	7.7 (0.9)	64 (8)	227 (20)
ω	0.02	0.01	0.22	0.28	0.46	
	KL	0.09	0.002	0.16	0.05	0.01
	lenis	66 (12)	338 (7)	4.7 (2.5)	49 (7.6)	152 (12)
bias + enhancement	aspirated	67 (19)	341 (10)	7.3 (0.9)	65 (9.6)	248 (17)
	ω	0	0.04	0.16	0.19	0.62
	KL	0.09	0.002	0.09	0.06	0.008
	lenis	65 (11)	338 (10)	5.5 (1)	48 (8)	170 (10)
target	aspirated	73 (15)	343 (12)	7.5 (1)	64 (9)	250 (11)
	ω	0.06	0.03	0.16	0.14	0.61
(cf. initial)	KL	0.16	0.002	0.12	0.06	0.008

Table 4.11: Comparison of means, standard deviations, cue weights, and KL divergences from three simulation scenarios with attested values estimated from modern Korean data.VOT = voice onset time; VLEN = vowel length; BA = burst amplitude (in dB); H_1-H_2 (in dB); F_0 (in Hz). KL divergence measured in bits.

4.4 General discussion

The simulation results presented above suggest that cases of (trans)phonologization in which promotion of a redundant cue is accompanied by decreased informativeness of a primary cue may be accounted for in a model where probabilistic enhancement is an adaptive response to bias threatening the primary cue to an existing contrast. This is not to say that phonologization must always be driven exclusively by bias, or that the presence of bias will invariably

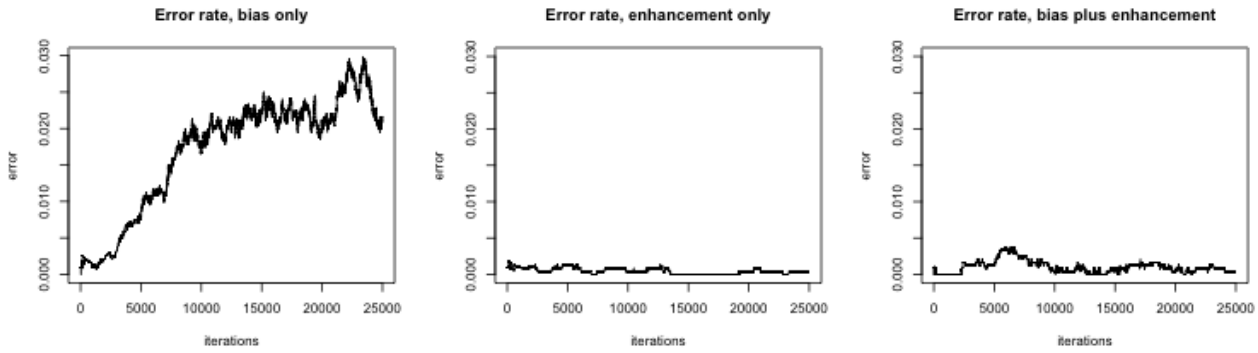

Figure 4.5: Comparison of contrast precision as measured by classification error rate at each simulation timestep for simulations reported in §4.3.1–4.3.3.

result in phonologization. To be sure, there are cases in which phonetic bias can condition the neutralization of the phonological contrast (see Chapter 5). Nevertheless, these results indicate that at least some cases of phonologization in which the relative balance of cue reliability has shifted may be the result of enhancement *in response to* a systemic production bias. As measured by KL divergence, the distributions resulting from the application of both enhancement and bias are more similar to the target distributions than those resulting from the application of only enhancement or only bias. The present account goes beyond the observation that a system biased against one cue will choose another to suggest that *which* cue takes over can be predicted with at least some degree of accuracy. Crucially, this account assumes that the speaker plays an active role in sound change, actively enhancing phonetic cues in order to accommodate the communicative needs of listeners.

The present account also suggest answers to the SELECTION and TRADING PROBLEMS laid out in Chapter 1. In this model, the selection of cues as potential targets of phonologization is probabilistically related to their distributionally-defined reliability ω. As a result, cues which are, all else being equal, more reliable are more likely to be enhanced by speakers. While it is not impossible for a relatively uninformative cue to become the primary indicator

of a contrast, it is highly unlikely, as the cumulative effect of incremental enhancements makes it less and less likely that the balance of cue reliability will shift significantly (barring additional bias factors which may affect their realization or perception, of course).

There are two ways to view the TRADING PROBLEM on the present account. One is that, if cue reliability is a normalized measure, an increase in the reliability of one cue will result in a decrease in reliability of all other cues (if reliability is taken to be purely distributionally-defined). The problem with this view is that it allows for scenarios in which a phonological contrast can be reliably differentiated along any number of individual cue dimensions which nonetheless of wildly varying levels of reliability. This fails to reflect the oft-observed empirical fact that phonologization of a once redundant cue tends to be accompanied by a loss of contrast along the previously primary cue dimension, in the sense that the contrast can no longer be reliably differentiated along that dimension alone. For this reason, I propose here that, at least for some cases (such as the Seoul Korean case), TRADING instead results from the *interaction* of one or more bias factors with the propensity for adaptive enhancement.

4.4.1 Bias factors

While the preceding simulation results illustrate a possible pathway by which F0 was phonologized in Seoul Korean, it also raises a number of questions and leaves unresolved a number of important issues. One of the most obvious questions concerns the nature of the bias factors. The model here says nothing about *why* a bias should be present; it is only able to predict the likely outcome given a bias of some magnitude. Specifying the full range and source of bias factors is of course an ongoing research effort in its own right (Ohala, 1981a *et seq.*; Blevins, 2004; Garrett and Johnson, to appear). Nevertheless, it is useful to have a highly explicit framework such as that presented above, with which the potential effects of *arbitrary* biases on the system of phonological contrasts can be explored.

In addition, the effects of different *degrees* of bias have yet to be fully explored within the current framework. How does the likelihood of phonologization vary when bias is implemented for different categories and different cues? What are the effects of intergenerational biases, as opposed to the intragenerational ones explored here? These are important questions, because as the model is currently specified, there is actually nothing privileged about bias itself: it is just one way to decrease contrast precision, which is what fundamentally modulates the likelihood of enhancement. Other factors which could decrease classifier accuracy, such as phonetic context, should be considered in future extensions of this model.

Related to the issue of bias is the role of the β exponent specifying the system-wide importance of the contrast. Along with ϵ, the measure of classifier accuracy, β can be used to tune both the rate at which enhancement will occur as well as the degree of phonetic exaggeration of means and restriction of variance. While the theoretical motivation for this parameter is the notion that some contrasts are more lexically or morphologically informative than others, its value in these simulations was not systematically manipulated. Ideally, its value could be derived by external principles, e.g. in the information-theoretic quantification of functional load proposed by Surendran and Niyogi (2003, 2006).

4.4.2 Cue relations and speaker control

A crucial assumption made in the present model is that cues are conditionally independent. While not wholly unmotivated (Clayards, 2008), such independence assumptions are almost certainly not warranted in every case: that is, there are some cues which are not independent by dint of physiological (articulatory or perceptual) factors (such as (perceived) vowel length and pitch, or amplitude reduction and formant bandwidth). An example of this is the fact that, in empirical studies of sound change in Korean, F0 actually *increased* for both categories as VOT became neutralized. If bias in VOT productions only affected one category (say, lenis stops), the F0 values of lenis stops might also be increasing because of a physiological

correlation of VOT and F0.

On the other hand, for some pairs of cues (closure duration and F1 at vowel onset, for instance) the independence assumption is probably warranted. One way to capture these dependencies would be to model categories using a multivariate normal distribution in a higher-dimensional space. Such a representation would allow for possibly complex covariance relations between different cue dimension; however, the sparse data space could result in poor parameter estimates (Toscano and McMurray, 2010). In terms of the general architecture proposed here, however, nothing prohibits moving from multiple, independent univariate GMMs to a single multivariate GMM representation.

This account also makes certain assumptions about the degree of speaker control over individual cues: in particular, that anything used as a perceptual cue is, in principle, amenable to active, targeted enhancement in speech production. Laboratory work on enhancement, however, suggests that speakers may employ a variety of methods to increase intelligibility of speech, and that the phonetics of 'clear speech' are highly complex. In several studies of the enhancement of stop contrasts, for instance, researchers have observed enhancement of VOT, but only for one member of a two-way contrast (Kessinger and Blumstein, 1997; Kang and Guion, 2008; Smiljanić and Bradlow, 2008). Kingston and colleagues (Kingston and Diehl, 1994; Kingston et al., 2008) have argued that cues are enhanced based on the degree to which they contribute to the perception of an INTEGRATED PERCEPTUAL PROPERTY (or IPP) which reinforces an existing phonological contrast. In the case of the [voice] contrast for initial stops, for example, cues with similar auditory properties, such as F1 and F0, are predicted to integrate, while cues such as closure duration and F0 are not, because they do not both contribute to the amount of low-frequency energy present near a stop consonant (Kingston et al., 2008). If cues are enhanced based on the degree to which they contribute to IPPs, cues such as closure duration should not be enhanced. While the core notion of probabilistic enhancement is not dependent on the assumption that speakers can

in principle exert active control over the realization of any and all phonetic cues, the model as implemented here makes different empirical predictions than previous accounts which deserve further investigation.

4.4.3 Phonologization, neutralization, and subphonemic reorganization

It is important to not dismiss cases like Korean as not constituting 'true' instances of phonologization on the grounds that subphonemic reorganization has not (yet) resulted in the emergence of a *new* phonological contrast. On such an understanding of the term, cases such as the emergence of vowel length as a cue to word-final obstruent voicing in English would similarly be rejected from consideration on the grounds that English does not distinguish minimal pairs solely on the basis of vowel length. Although the question of how and when acoustic-phonetic cues become phonological features – or if this is even the proper analysis of the situation – is undoubtably important, so too is understanding how the stage for such a phase transition is set. A framework for modeling shifts in phonetic cue weights such as that presented here is critical for understanding why such phase transitions take place in some languages and not others, and under certain distributional circumstances and not others.

Although the focus in this chapter has been on the internal dynamics of contrast preservation through transphonologization, one might reasonably ask how (or if) the present model handles cases of neutralization. Neutralization is possible within the present framework and is in fact predicted under certain circumstances: if a bias factor is strong enough (and/or the enhancement constant β is large enough, implying a contrast with a low functional load), the prediction is complete neutralization the contrast. This is because even if probabilistic enhancement is applied in such cases, it may not be strong enough to counteract the effects of bias. This is likely to be the case when a covert contrast does *not* already exist, i.e. when a contrast is maintained essentially by a single cue dimension. By allowing the number of mixture components to vary, a distinction between two categories is predicted to collapse

when the optimal model contains just a single mixture component. Chapter 5 discuss this type of scenario in greater detail.

The nature of iterated learning framework proposed here – in which an agent's internal state can change over time – assumes that intragenerational sound change is possible. While Harrington, Palethorpe, and Watson (2000a,b) show that the phonetic targets of older members of a speech community may subconciously shift in response to broader changes taking place in the community, it is almost certainly true that phonetic change also occurs during intergenerational transmission, either as a function of learning and induction mechanisms (S. Kirby, 1999; S. Kirby and Hurford, 2002; McMurray et al., 2009) and/or as a result of listener misperceptions (Ohala, 1981a; Blevins and Garrett, 1998; Blevins, 2004). Indeed, the main source of the data used in the simulations reported here, Kang and Guion (2008), is an apparent time study, suggesting that the the production targets for speakers may not change significantly over the course of that speaker's lifetime. Thus, it is worth considering the possibility that, in addition to (or in place of) enhancement as a response to systematic bias, phonologization may result from a dynamic shift in the course of intergenerational transmission (Niyogi, 2006). In order to test this, the simulation architecture will need to be extended to cover multiple agents in several generations. Similarly, the architecture should be expanded to allow for lexical factors such as frequency and neighborhood density to influence the dynamics of sound change (Wang et al., 2004; Wedel, 2006; Martin, 2007; Blevins and Wedel, 2009).

4.5 Conclusion

This chapter has explored the role of probabilistic enhancement in phonologization, introducing an iterated learning framework for simulating the production and perception of phonological contrasts along with the results of a number of simulations making use of this framework. The results of applying a strategy of probabilistic enhancement in addition to a phonetic bias

factor were found to most closely mirror an attested instance of phonologization in which a systemic production bias exposed an existing covert contrast. This strategy addresses two of the challenges faced by a phonologization model of sound change, dubbed the SELECTION PROBLEM and the TRADING PROBLEM. The answer proposed to the SELECTION PROBLEM ('which cue(s) are targeted for phonologization?') is that the enhancement of cues is a probabilistic function of their reliability. Thus, a cue which may be informative (and therefore targeted for enhancement) in one language may be ignored in another. The answer proposed to the TRADING PROBLEM ('why is phonologization accompanied by neutralization?') is that attention to cues is relative: if the cue space is perturbed in such a way as to reduce the reliability of one cue dimension, others will be enhanced proportionally. Depending on the nature and degree of enhancement, the result can set the stage for a reorganization of the system of phonological contrasts.

CHAPTER 5

PHONETIC CATEGORY RESTRUCTURING

The preceding chapter demonstrated how loss of contrast precision induced by phonetic bias might drive the transphonologization of a previously redundant phonetic cue. Empirically speaking, however, contrasts do not always survive a bias-induced loss of precision. The collapse of a contrast between previously distinct categories is called MERGER; a contrast which does not obtain in a certain context is said to be NEUTRALIZED in that context. In many dialects of American English, for example, it has been claimed that the contrast between /t/ and /d/ is neutralized to [ɾ] when followed by an unstressed vowel, leading to homophony between pairs such as *metal–medal* and *cuttle–cuddle* (Giegerich, 1992). In Dutch, the voicing contrast between obstruents in word-final position has been argued to result in homophony between word pairs such as those in Table 5.1 (Lahiri et al., 1987).

voiceless			voiced		
baat	/bat/	'benefit'	*baad*	/bad/	'bathe-1sg'
noot	/not/	'nut'	*nood*	/nod/	'necessity'
voet	/vut/	'foot'	*voed*	/vud/	'feed-1sg'

Table 5.1: Dutch minimal pairs differing in underlying voicing of final obstruent.

In terms of the mixture model of phonetic categories, neutralization or merger may be thought of as a reduction in the total number of category labels. Up to this point, we have been assuming that the set of phonetic category labels is given in advance. In order to understand how the number of labels might change (grow or shrink), it is necessary to consider how the label set might be inferred in the first place. This chapter will introduce a method for inducing both the number of components of a mixture model as well as their parameterization, and illustrate how and under what conditions it predicts an ADAPTIVE RESTRUCTURING of the phonetic category system.

5.1 Unsupervised induction of phonetic categories

The previous chapters have focused on classifying vectors of acoustic-phonetic cues: that is, selecting an appropriate label for a new observation, given an extant set of category labels. However, it is also worth considering the issue of where the category labels themselves come from, and how they might be inferred from unlabeled data. Work on the acquisition of phonological contrasts has demonstrated that while human infants are initially quite sensitive to a wide variety of speech sound contrasts, this ability quickly becomes lost in the course of early development (Werker, Gilbert, Humphrey, and Tees, 1981; Best, McRoberts, and Sithole, 1988; Bosch and Sebastián-Gallés, 2003; Kuhl et al., 2006). This suggests that phonetic category structures may be formed and solidified on the basis of acoustic input alone, formed in response to the distribution of acoustic cues in the input. This has inspired a growing number of researchers to apply computational techniques of clustering and pattern-recognition to the problem of phonological category induction in considering how human infants might learn the phonological categories or words from the ambient language without explicit instruction (de Boer and Kuhl, 2003; Lin, 2005; Vallabha et al., 2007; Feldman et al., 2009; McMurray et al., 2009).

Here, the same general approach is applied to the study of category restructuring in adults. Adult listeners are constantly being exposed to new data, and so constantly have the opportunity to revise their set of category labels. Given the assumptions about memory decay implemented in Chapters 3 and 4, inferences about the structure of the cue space are predicted to change over time in accordance with linguistic experience. The goal of the present chapter is to better understand how changes in the structure of the cue space impact the adaptive restructuring of the category label space. First, I will discuss the technical aspects of the clustering procedure, and then illustrate the predictions it makes regarding category restructuring with reference to data from Korean and Dutch.

5.2 Model-based clustering

MODEL-BASED CLUSTERING (Fraley and Raftery, 2002, 2007) is a powerful unsupervised learning technique which allows for information or assumptions about the underlying distribution of the observation data to be modeled directly, usually in the form of a finite mixture model of the type introduced in Chapter 2, where each cluster (component) is assumed to have its own likelihood and underlying probability distribution. The clustering procedure described here builds on the GAUSSIAN MIXTURE MODEL (GMM) of speech categories. Recall that in a GMM, an observed data point $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)$ is assumed to be generated by a mixture model with density

$$f(\mathbf{x}; \theta) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x} | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{5.1}$$

where π_k is the kth component weight and $\theta = (\theta_1, \ldots, \theta_K) = ((\pi_1, \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1), \ldots, (\pi_K, \boldsymbol{\mu}_K, \boldsymbol{\Sigma}_K))$ is a $K(D+2)$-parameter structure containing the component weights π_k as well as the mean vectors $\boldsymbol{\mu}_k$ and covariance matrices $\boldsymbol{\Sigma}_k$ of the D-variate component Gaussian densities

$$\mathcal{N}(\mathbf{x} | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = \frac{1}{(2\pi)^{D/2} |\boldsymbol{\Sigma}_k|^{1/2}} \exp\left\{ -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\mathbf{x} - \boldsymbol{\mu}_k) \right\} \tag{5.2}$$

where the component weights $\pi_1, \ldots, \pi_K$ sum to 1.

As mentioned in Chapter 2, expectation maximization (EM) can be used to obtain ML estimates of GMM parameters. Starting from an initial guess about θ, the EM algorithm alternates between computing a probability distribution over completions of missing data given the current model and then re-estimating the model parameters using these completions.

While an iterative process such as EM is useful for finding the parameters θ of a GMM, in a truly unsupervised learning scenario there remains an additional problem: how to determine the best value of K. The criterion of maximum likelihood in and of itself is of little use here, as maximum likelihood may be achieved by associating each observation with its own

Gaussian, so that the final model has as many Gaussians as it does data points. What is needed is some means of constraining the number of components K in the model, a process termed REGULARIZATION (McLachlan and Peel, 2000; Hastie, Tibshirani, and Friedman, 2008).

In practice, regularization is often achieved through a process called CROSS-VALIDATION. In cross-validation, a separate set of data not seen in training is used to evaluate the accuracy of some other number of models. An advantage of cross-validation, in addition to the fact that it often achieves good estimates in practice, is that it makes no assumptions about either the underlying model (i.e., is it parametric or non-parametric) or the task (it may be applied to problems of regression or density estimation in the same fashion as classification and clustering). The primary drawback of using cross-validation is that it requires a gold standard or ground truth in order to determine a meaningful measure of error. In the unsupervised learning of linguistic units such as phonemes or morphemes, these standards are usually constructed by hand. For instance, Singh, Raj, and Stern (2000) use the phoneme sequences from the CMUdict pronunciation dictionary (CMUdict, 1998) as the standard against which to test an unsupervised model of subword unit acquisition; the clustering experiments of Lin (2005), along with most automatic phone recognition systems, are evaluated using the hand-labeled TIMIT database (Garofalo et al., 1993). The optimality of the best solution is thus constrained by one's conviction in the appropriateness of the chosen gold standard.

Approaches which seek an optimal trade-off between data likelihood and model complexity represent a different type of regularization. Three of the most well-known criteria of this type are the AKAIKE INFORMATION CRITERION (AIC: Akaike, 1974), the BAYESIAN INFORMATION CRITERION (BIC: Schwarz, 1978; Fraley and Raftery, 2007), and the MINIMUM DESCRIPTION LENGTH principle (Rissanen, 1978; Grünwald, 2007). All of these criteria are closely related and contain two core terms, one measuring goodness of fit and one measuring model complexity.

Both the AIC and the BIC are derived from the log-likelihood of the observation data and the number of parameters of a proposed model. Given a series of N independent, identically distributed D-dimensional observations $\mathbf{X} = \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N$, let $\mathcal{L}$ be the maximized log-likelihood of a GMM with K D-dimensional components characterized by parameters θ, and let $\mathcal{Q}$ stand for the number of independent parameters in that model:

$$\mathcal{L} = \ln P(\mathbf{X}|\theta_{\max}) \tag{5.3}$$

$$\mathcal{Q} = K(D + D(D+1)/2) + K - 1 \tag{5.4}$$

The AIC is then defined as

$$AIC = -2\mathcal{L} + 2\mathcal{Q} \tag{5.5}$$

and the BIC is defined as

$$BIC = -2\mathcal{L} + \ln(N)\mathcal{Q} \tag{5.6}$$

In both calculations, fit is measured negatively by $-2\mathcal{L}$; the larger the value, the worse the fit. Complexity is measured positively, either by $2\mathcal{Q}$ (AIC) or $\ln(N)\mathcal{Q}$ (BIC). Given two models fit on the same data, the model with the smaller AIC or BIC value is considered superior in terms of the fit-complexity trade-off. As N increases, the BIC tends to favor simpler models than the AIC, due to stronger penalty imposed by the BIC for increased model parameters.

The Minimum Description Length (MDL) approach to model selection penalizes the complexity of a model based on its ENCODING LENGTH. Given an encoding of length ℓ bits representing a model with parameters θ, the MDL principle says to pick the model with the minimum DESCRIPTION LENGTH $\mathcal{D}$:

$$\mathcal{D}(\mathbf{X}, \theta) = \ln \mathcal{L} + \ell(\theta) \tag{5.7}$$

Here, the code length ℓ plays the role of the penalized likelihood penalties in AIC and BIC. MDL has connections to both the AIC (in that KL divergence plays a crucial role in both approaches) as well as to the BIC (which can be interpreted as a special case of MDL under certain circumstances: see Hansen and Yu, 1999, 2001).

Some previous researchers who fit mixture models to speech data using EM, such as de Boer and Kuhl (2003) and Lin (2005), assumed a fixed K and did not address the issue of regularization. Analytic criteria such as the BIC build on these works by offering a theoretically principled approach to model selection. A different approach is taken by researchers such as Vallabha et al. (2007) and McMurray et al. (2009), who employed a 'winner-takes-all' approach to model selection: starting from some suitably large K, the mixture parameters are updated after each input, but the component weight is updated only for the most likely component. This type of model has the important property that it can learn online, as opposed to the batch learning of BIC-based model selection, but requires setting a threshold below which categories are eliminated from further consideration.

Feldman et al. (2009) take an explicitly non-parametric Bayesian approach to the problem of phonetic category induction, allowing a potentially infinite number of components K but imposing a prior that is biased toward mixtures with smaller numbers of categories (Rasmussen, 2000; Teh, Jordan, Beal, and Blei, 2006). While conceptually and computationally elegant, this approach has the potential drawback that the degree of bias is a free parameter of the model.

In what follows, we will use the BIC because of its computational tractability, because it imposes a stronger penalty on model complexity than the AIC, and because it does not require a gold standard for comparison. Since all approaches to regularization embody the notion that the communicative code should be maximally robust while remaining as compact

as possible, the choice of one method over another is in practice largely driven by trying multiple approaches and choosing the one with the best performance; from a conceptual standpoint, there may not be a compelling theoretical reason to choose one over another. That having been said, it must be recognized that the results using a different form of regularization could potentially be quite different, but an explicit comparison between the various approaches will not be attempted here. In addition, no claims are made regarding the accuracy of the BIC or any other methods at modeling human performance data (but see Pothos and Chater, 2002; Pothos and Close, 2008; Pothos and Bailey, 2009).

5.3 Category restructuring as model selection

In sound change, neutralization (merger) is said to occur when the productions of two (or more) phonological categories become acoustically (and perhaps articulatorily) identical, presumably because at this stage, listeners can no longer perceptually distinguish between categories. The result of a merger can be either STRUCTURE-PRESERVING, if listeners analyze productions of one extant category as intended instances of another) or STRUCTURE-CREATING, if listeners restructure the label space in such a way that observations that previously received distinct labels are assigned a new label on the basis of a cue distribution that does not conform to either the distributions of the original categories.

Given the mixture model of categories developed in Chapter 2 combined with the regularization and model selection techniques discussed above, this theory might be implemented in the simulation framework described in §3.1 by having agents compute the likelihood and number of parameters of models with some range of components (and potentially with different parameter structures) at the end of each simulation iteration. If the BIC-optimal number of categories differs from that of the previous iteration, the listener re-labels the contents of its memory accordingly.

Of the three simulation scenarios considered in Chapter 4, two (enhancement only and

enhancement + bias) resulted in outcomes in which the Korean lenis/aspirate category contrast was well-separated along several acoustic dimension, while the outcome of §4.3.2, in which phonetic bias was implemented but adaptive enhancement was not, was argued to be the most likely candidate for neutralization. As such, we will first use the output of this simulation to test the predictions of model-based clustering using the BIC. The internal state of the agents after 25,000 simulation iterations are replicated in Figure 5.1 and Table 5.2.

Figure 5.1: Cue distributions after 25,000 iterations for lenis /p/and aspirated /p$^{\text{h}}$/ stops, VOT bias-only simulation condition. Captions give cue reliability ω.

Category	VOT	VLEN	H_1-H_2	BA	F_0
lenis	65 (11)	340 (8)	6.3 (1.8)	48 (7)	162 (12)
aspirated	65 (16)	340 (9)	7.7 (0.9)	64 (8)	227 (20)
ω	0.02	0.01	0.22	0.28	0.46

Table 5.2: Means, standard deviations, and cue weights after 25,000 iterations of a bias-only simulation scenario discussed in Chapter 4. VOT = voice onset time; VLEN = vowel length; BA = burst amplitude (in dB); $H_1 - H_2$ (in dB); F_0 (in Hz).

To determine if neutralization of /p/ and /p$^{\text{h}}$/ would be predicted under these conditions, we must consider several scenarios. First, we need to limit how many components K_{max} we would be willing to allow a model to have. For present purposes, we will consider models with K in the range 1 to 5. Five different models were fit to the contents of one agent's memory after 25,000 simulation iterations.[1] Given the α memory weight value used in these simulations, the agent's memory contained 2,408 exemplars at the end of the simulation (1,167 labeled /p/ and 1,241 labeled /p$^{\text{h}}$/), so $N = 2408$ in this example.

1. The volume, shape and orientation of the covariance matrices were allowed to vary in model fitting; only the best (i.e. BIC-optimal) results are reported here. See Fraley and Raftery (2006, 2007) for details.

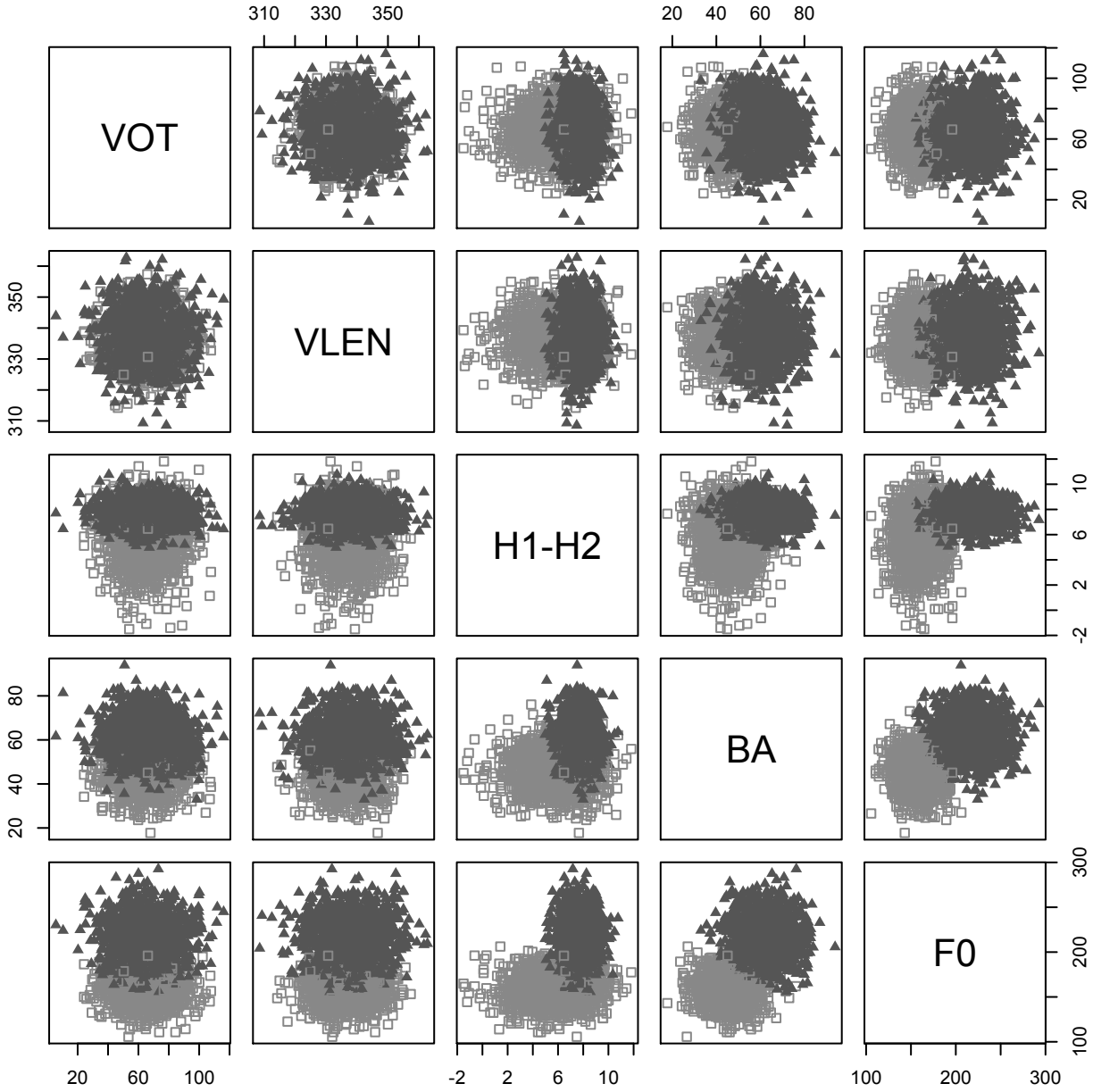

Figure 5.2: Symmetric pair plot showing BIC-optimal classification of contents of agent memory after 25,000 simulation iterations in which bias was applied to VOT productions but enhancement was not implemented. Gray squares show predicted instances of lenis /p/, black triangles aspirated /p^h/ stops.

K	*BIC*	*error*
1	84825	0.49
2	**84096**	**0.02**
3	84175	0.02
4	84250	0.03
5	84316	0.05

Table 5.3: BIC scores and classification error rates for models of 1-5 components, VOT bias only condition. Optimal solution given in bold. Bayes error of an optimal two-component classifier = 0.02. Error rates correspond to a minimum error mapping between the predicted classification and the ground truth.

The results are shown in Figure 5.2 and Table 5.3. Despite the high degree of overlap along almost all of the cue dimensions, the BIC-optimal model in this case still has two categories, not one as would be predicted if neutralization was the expected result. The pair plots shown in Figure 5.2, which give the predicted categorization of the contents of the agent's memory at the end of the simulation, suggest that the resulting model is still quite accurate at distinguishing between instances of lenis and aspirated stops, and indeed the resulting accuracy of the optimal two-component model confirm this, achieving the Bayes error rate of a two-component classifier.

This result seems to suggest that, if along just a single acoustic-phonetic dimension, even considerable overlap may not be sufficient to induce a restructuring of phonetic categories. It is then worth considering the question of just how much overlap between cue dimensions is necessary to predict neutralization. To test this, a similar bias-only Korean simulation was conducted in which the bias factor affected *all* cues simultaneously. As in previous simulations, this bias affected both the difference between categories means and the variances, to make productions of one category more similar to instances of the other (or less dissimilar from) with probability and to a degree relative to the cue's reliability and the precision of the contrast, as in Chapter 3.

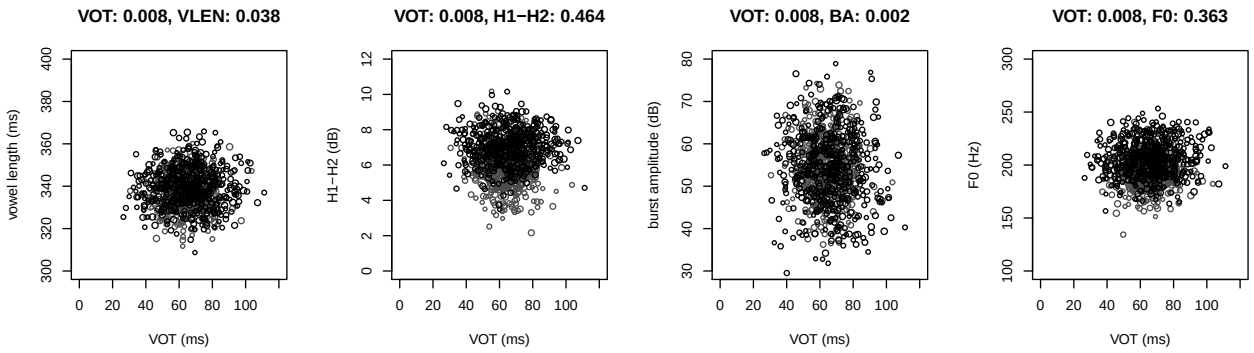

Figure 5.3: Cue distributions after 25,000 iterations for lenis /p/and aspirated /p^h/ stops, across-the-board leniting bias simulation condition. Captions give cue reliability ω.

Category	VOT	VLEN	H_1-H_2	BA	F_0
lenis	65 (11)	337 (8)	6 (1.2)	58 (8)	192 (13)
aspirated	67 (15)	340 (9)	7 (1)	54 (9)	207 (17)
ω	0.05	0.12	0.31	0.16	0.35

Table 5.4: Means, standard deviations, and cue weights after 25,000 iterations of a bias-only simulation scenario in which all five cues (including F0) are subject to leniting bias. VOT = voice onset time; VLEN = vowel length; BA = burst amplitude (in dB); $H_1 - H_2$ (in dB); F_0 (in Hz).

Given the α memory weight value used in this simulation, the agent's memory again contained 2,408 exemplars at the end of the simulation (1,209 bearing the label /p/ and 1,199 bearing the label /p^h/). The resulting distributions are given in Figure 5.3 and Table 5.4; the results of model-based clustering are shown in Table 5.5. In spite of almost complete overlap along all cue dimensions, the optimal model on the BIC metric is once again a model with two components: in other words, neutralization is still not the predicted optimal outcome.

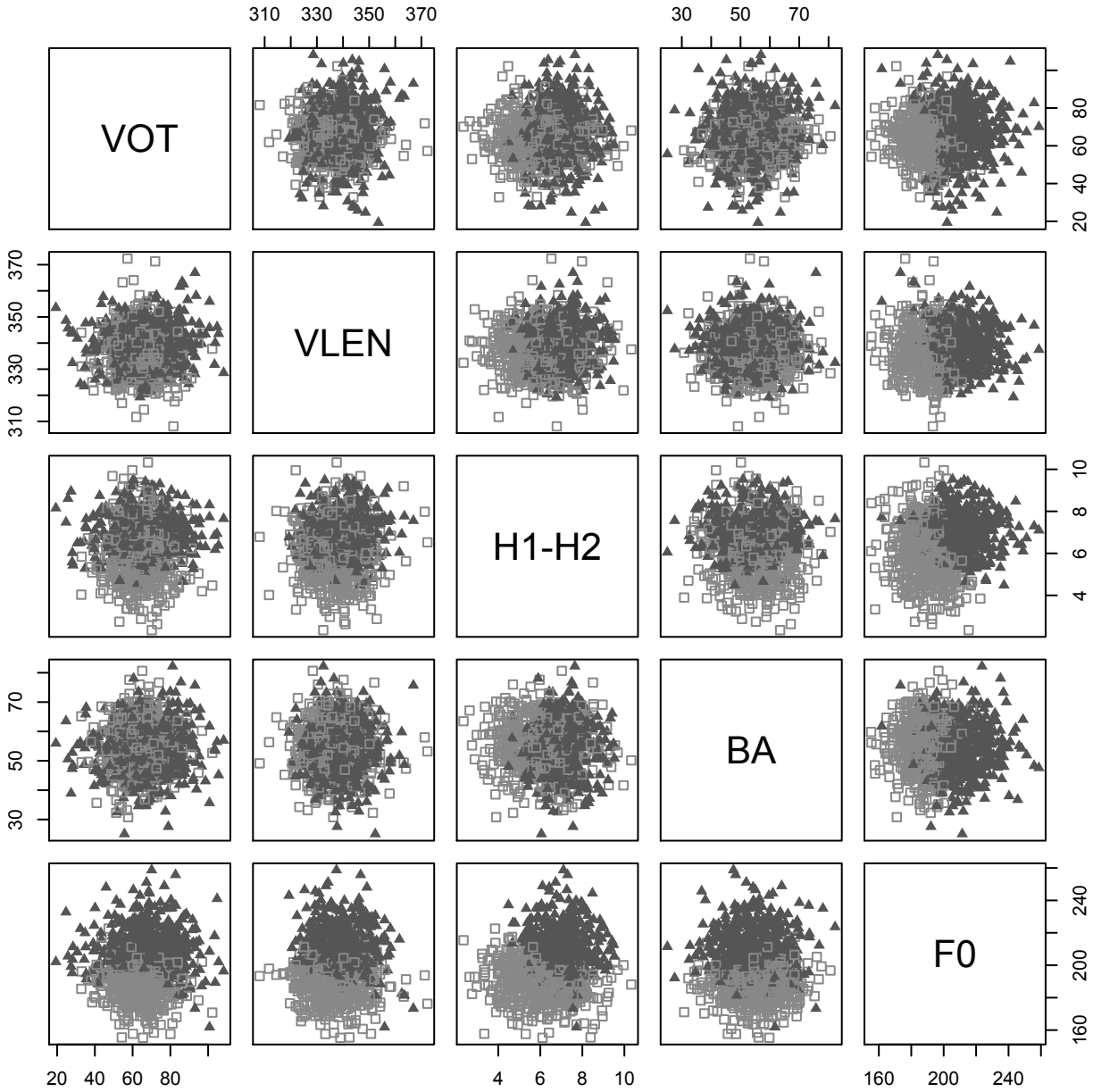

Figure 5.4: Symmetric pair plot showing optimal classification of contents of agent memory after 25,000 simulation iterations in which bias was applied to production of all cues. Gray squares show predicted instances of lenis /p/, black triangles aspirated /p^h/ stops.

K	*BIC*	*error*
1	34072	0.50
2	**34021**	**0.35**
3	34058	0.40
4	34084	0.39
5	34117	0.37

Table 5.5: BIC scores and classification error rates for models of 1-5 components, pure lenition. Bayes error rate of an optimal two-component classifier = 0.23. Error rates correspond to a minimum error mapping between the predicted classification and the ground truth.

5.3.1 Separability in high dimensionality

Despite considerable overlap along multiple dimensions, neutralization is not predicted in either of the Korean scenarios discussed above. This might be taken as evidence that BIC-based model selection is an inappropriate means of predicting whether or not a contrast will neutralize. However, it may also be the case that, although the categories do not appear well-separated when considering any individual cue dimension, they are well-separated in a higher-dimensional space.

To see how this is trivially true, consider the data in Figure 5.5, which is based on a subset of data from an acoustic study of the voiced Dutch stops /b/ and /d/ (Smits, ten Bosch, and Collier, 1996a,b). Figure 5.5(A) shows F2 frequency at voicing onset of a male Dutch talker for tokens of /bV/ and /dV/, where V = {/a i y u/}. Given just this single cue dimension, an optimal classifier (boundary indicated by the dashed line in Figure 5.5A) would misclassify 25% of the tokens, but when provided with information about F2 at the nucleus of the following vowel (Figure 5.5B), the classification error rate drops to 12.5%.

Smits (1996) uses this example to illustrate several important points. First, the more acoustic dimensions the classifier has access to, the lower the classifier's potential (if not actual) classification error rate. Thus, large within-category variability for a given cue may only be problematic inasmuch as it is the only cue available on which to make a categoriza-

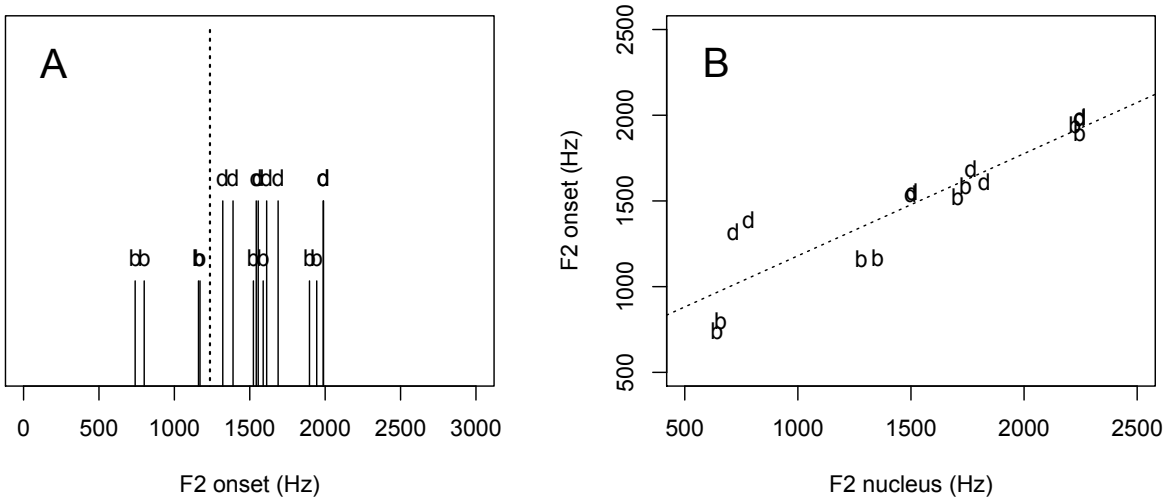

Figure 5.5: Instances of Dutch /b/ and /d/ in onset position in (A) one and (B) two acoustic dimensions. Dashed lines give the optimal class boundaries. Adapted from Smits (1996).

tion decision. Second, utterances that are acoustically very different along several phonetic dimensions may receive the same label, while utterances that are acoustically very similar may receive different labels, as can be seen in Figure 5.5. Finally, the potential contribution of a cue in distinguishing a contrast can only be established given the total set of cues that are deemed relevant for perception of the contrast (Port and Crawford, 1989).

All of these observations underscore the point that separability cannot be predicted from consideration of a single acoustic dimension. This may seem a trivial observation from a statistical standpoint, but as Smits notes,

> it does steer phonetic problem formulations away from classical questions like "what is the best or most disambiguating cue" towards issues like "what is the cue space dimensionality" and "what set of cues leads to good or human-like classification." (1996)

Note that from the standpoint of the statistical classifiers discussed so far, all cue dimensions are potentially of equal reliability: it is their distributional properties alone that give rise to the observed differences. As noted in §2.4.4, however, cue informativeness can be modulated by attention and context (Gordon et al., 1993; Francis, Kaganovich, and Driscoll-Huber, 2008) as well as by task (Iverson, Hazan, and Bannister, 2005; Holt and Lotto, 2006), and the predictions about diagnosticity vary with theories of how cues are integrated in perception (Garner, 1974; Gordon et al., 1993; Francis et al., 2008; Kingston et al., 2008; Toscano and McMurray, 2010). A complete solution to the RESTRUCTURING PROBLEM will need to provide a dynamic mechanism to address these factors.

5.4 The effects of cue availability on category restructuring

Section 5.3 demonstrated how the separability of phonetic categories may be heavily dependent on cue availability. It follows that cue availability may also impact the clustering of the acoustic-phonetic space. This section reports on two types of experiments that were conducted to explore the effects of cue availability on category restructuring. First, how does varying the number and combination of cue dimensions available to a statistical learner influence the (BIC-)*optimality* of a given solution? Second, how *typical* is any individual solution given a set of model parameters and a (sub)set of available cues?

5.4.1 Series 1: Optimality

The first set of experiments varied the number and type of cue dimensions made available to the learner in order to determine the predictions made by model-based clustering using the BIC about the restructuring of phonetic categories. Since model-based clustering can only usefully compare models fit to a single set of observations, the experiments in Series 1 fit a series of GMMs to a single set of $N = 500$ 5-dimensional observation vectors generated from a Gaussian mixture with the parameters given in Table 5.3 (the maximum likelihood parameter

estimates based on agent memories after 25,000 simulation iterations of a full-spectrum bias with no enhancement) using the EM-based estimator implemented in Fraley and Raftery (2006). The BIC scores of models with access to all possible unordered combinations of cue dimensions were then compared. The model-fitting procedure and observation data were identical in all experiments; only the number of cue vector columns available to the learner changed.

An overview of the results for a single set of simulations is given in Table 5.6. Since models of $K > 2$ were never BIC-optimal, BICs and classification error rates for models with $K = 3$ components are provided for comparison only.

The first thing that is clear from Table 5.6 is that (BIC-)optimality is not the same as accuracy: in roughly half the models tested, a simpler model (with a single component) was preferred to a more complex model (with two or more components), despite the fact that the single-component models all necessarily perform at chance.[2] Second, the presence or absence of any *single* cue dimension as an available column in the observation vector does not appear to determine the optimal number of categories, regardless of the (relative) reliability of that dimension (i.e., the degree to which it covaries with the underlying category distinction). However, note that when two of the most reliable dimensions (H_1-H_2 and F_0) were included, two categories were always BIC-optimal. Finally, note that the best accuracy was not achieved by a model with access to all five cue dimensions: the most accurate model among those considered here had access to just three (VOT, H_1-H_2, and F_0).[3]

2. In some cases, accuracy may also be improved by considering different covariance structures, but these models are similarly suboptimal based on the BIC.

3. Note that the accuracy of a classifier containing multiple cues (say, H_1-H_2 and BA) is not always monotonically decreasing relative to a classifier containing a proper subset of those cues (say, H_1-H_2). This is because only the BIC-optimal model accuracy is shown here, which may differ from the model with the maximum log-likelihood. Even though the same set of observations is used in all cases, the log-likelihood of those observations differs depending on how many vector columns are exposed to the learner.

cue dimension					$K = 1$		$K = 2$		$K = 3$	
d_1	d_2	d_3	d_4	d_5	*BIC*	*err*	*BIC*	*err*	*BIC*	*err*
VOT					**8004**	**0.50**	8018	0.44	8031	0.45
VLEN					**7263**	**0.50**	7278	0.42	7282	0.45
H_1-H_2					**3181**	**0.50**	3184	0.34	3198	0.41
BA					**7070**	**0.50**	7084	0.49	7098	0.47
F_0					8554	0.50	**8546**	**0.42**	8558	0.40
VOT	VLEN				**15267**	**0.50**	15289	0.42	15309	0.46
VOT	H_1-H_2				11183	0.50	**11180**	**0.32**	11190	0.50
VOT	BA				**15074**	**0.50**	15093	0.48	15105	0.45
VOT	F_0				16558	0.50	**16553**	**0.32**	16570	0.33
VLEN	H_1-H_2				**10444**	**0.50**	10460	0.32	10480	0.42
VLEN	BA				**14333**	**0.50**	14347	0.45	14365	0.45
VLEN	F_0				15810	0.50	**15802**	**0.33**	15821	0.36
H_1-H_2	BA				**10251**	**0.5**	10271	0.47	10293	0.45
H_1-H_2	F_0				11704	0.50	**11697**	**0.32**	11703	0.34
BA	F_0				**15625**	**0.50**	15631	0.32	15641	0.38
VOT	VLEN	H_1-H_2			18447	0.50	**18445**	**0.39**	18462	0.46
VOT	VLEN	BA			**22337**	**0.50**	22363	0.50	22385	0.45
VOT	VLEN	F_0			23821	0.50	**23810**	**0.38**	23836	0.45
VOT	H_1-H_2	BA			**18255**	**0.50**	18264	0.32	18289	0.39
VOT	H_1-H_2	F_0			19712	0.50	**19689**	**0.25**	19727	0.30
VOT	BA	F_0			23628	0.50	**23630**	**0.31**	23657	0.43
VLEN	H_1-H_2	BA			**17514**	**0.50**	17540	0.48	17565	0.49
VLEN	H_1-H_2	F_0			18964	0.50	**18955**	**0.27**	18986	0.32
H_1-H_2	BA	F_0			18786	0.50	**18780**	**0.27**	18789	0.42
VOT	VLEN	H_1-H_2	BA		**25518**	**0.50**	25521	0.33	25547	0.45
VOT	VLEN	H_1-H_2	F_0		26979	0.50	**26941**	**0.28**	26981	0.34
VOT	VLEN	BA	F_0		30891	0.50	**30886**	**0.40**	30919	0.42
VOT	H_1-H_2	BA	F_0		26801	0.50	**26764**	**0.28**	26790	0.31
VLEN	H_1-H_2	BA	F_0		26051	0.50	**26036**	**0.36**	26048	0.40
VOT	VLEN	H_1-H_2	BA	F_0	34072	0.50	**34021**	**0.35**	34058	0.40

Table 5.6: BIC scores and error rates for models in 2–5 dimensions. K = number of categories (components); columns show the cue dimensions made available in the observation data. Bold items indicate the optimal solutions. BIC values rounded to nearest integer value. Bayes optimal error rate for a two-component model: 0.226 (see §2.4.5).

cue dimension							
d_1	d_2	d_3	d_4	d_5	$K = 1$	$K = 2$	$K = 3$
VOT					**1.00**		
VLEN					**0.99**	0.01	
H_1-H_2					**0.90**	0.10	
BA					**1.00**		
F_0					**0.64**	0.36	
VOT	VLEN				**1.00**		
VOT	H_1-H_2				**0.80**	0.20	
VOT	BA				**1.00**		
VOT	F_0				0.41	**0.59**	
VLEN	H_1-H_2				**0.94**	0.06	
VLEN	BA				**0.99**	0.01	
VLEN	F_0				**0.57**	0.43	
H_1-H_2	BA				**1.00**		
H_1-H_2	F_0				0.35	**0.65**	
BA	F_0				**0.85**	0.15	
VOT	VLEN	H_1-H_2			**0.86**	0.14	
VOT	VLEN	BA			**1.00**		
VOT	VLEN	F_0			0.42	**0.58**	
VOT	H_1-H_2	BA			**1.00**		
VOT	H_1-H_2	F_0			0.42	**0.58**	
VOT	BA	F_0			**0.79**	0.21	
VLEN	H_1-H_2	BA			**1.00**		
VLEN	H_1-H_2	F_0			0.13	**0.87**	
H_1-H_2	BA	F_0			0.29	**0.71**	
VOT	VLEN	H_1-H_2	BA		**0.90**	0.10	
VOT	VLEN	H_1-H_2	F_0		0.02	**0.98**	
VOT	VLEN	BA	F_0		**0.60**	0.39	0.01
VOT	H_1-H_2	BA	F_0		0.10	**0.90**	
VLEN	H_1-H_2	BA	F_0		0.10	**0.90**	
VOT	VLEN	H_1-H_2	BA	F_0	0.04	**0.96**	

Table 5.7: Proportion of BIC-optimal category solutions for Korean data in terms of percentage of 1,000 fits. Most-typical (≥ 0.50) solution percentages given in bold.

5.4.2 Series 2: Typicality

Since the BIC is computed for models fit to a given set of observations, different sets of observations generated from the same distributional parameters could potentially lead to different solutions being BIC-optimal.[4] This means that the contents of a particular agent's memory at some particular point during a simulation might lead it to posit more or fewer categories, but that this could vary by individual. In order to assess the typicality of the solutions shown in Table 5.6, the experiments described in Series 1 were repeated 1,000 times for each combination of cue vectors, with a new set of $N = 500$ observations generated from the parameters given in Table 5.4 each time.

The results are given in Table 5.7. Over half the possible combinations of cue dimensions are characterized by some type of variability. However, there were no cases in which a two-component solution was always preferred (although several combinations tended to heavily favor two-component solutions, such as four-dimensional combinations containing both F_0 and $\mathrm{H}_1 - \mathrm{H}_2$, as well as when all five columns were present in the observation data). Three-component solutions were almost never optimal, except in a few instances for a single cue combination.

5.4.3 Discussion

The experiments described above suggest that model selection based on the BIC, a particular method of adjudicating between model fit and model complexity, may be an appropriate means of modeling and predicting phonetic category restructuring, but that the results depend heavily on the number, combination, and distribution of the cue dimensions used as input. Thus, on its own, the method simply provides a way to determine if a contrast *could* be learned from data in an unsupervised fashion, but does not establish that contrasts are

4. Observations sets of different size could similarly impact BIC-optimality; this factor was not systematically manipulated here, but is the subject of ongoing investigation.

in fact learned by such a method.

The cues to the lenis/aspirate contrast in the second set of Korean simulation results used as input to the experiments were highly overlapping in all five dimensions. This highlights the fact that the learnability or recoverability of a contrast may not always be obvious from the consideration of any single cue dimension in isolation. That is to say, even if a contrast does not appear to be separable along any individual cue dimension, it may still be separable when multiple dimensions are considered in concert. Empirically, one might then ask how many dimensions, and what combination, are necessary to achieve learnability.

The results of the replications in Series 2 demonstrate that even in cases where a contrast is relatively (im)precise, its fate cannot be predicted with absolute certainty; the outcome depends on the particular data set used for parameter estimation. In terms of the agent-based simulations conducted in the previous chapter, since different agents will have slightly different experiences, the contents of their memories will differ accordingly, as will the likelihood that they will posit changes to the overall structure of phonetic categories at a given point in time. If this is an accurate representation of the variation present in a speech community, it predicts that, empirically speaking, we should expect to find situations in which one a contrast is neutralized for some speakers but not for others. In the following section we will turn our attention to just such a case.

5.5 Covert contrast: the case of Dutch final devoicing

One empirical domain in which individual variation of this sort has been reported are cases of so-called INCOMPLETE NEUTRALIZATION or COVERT CONTRAST (Hewlett, 1988; Scobbie et al., 2000; Yu, 2007), also referred to in the sociohistorical tradition as NEAR MERGERS (Labov et al., 1972; Labov, Karen, and Miller, 1991). These are instances in which contrasts which appear to be neutralized to the naked ear of a linguist may in fact be distinguished by subtle yet statistically significant differences in production and perception. Reports of

incomplete neutralization and near merger often note that these changes tend not to be uniformly distributed throughout a population. That is to say, some members of population may still distinguish a contrast in production and/or perception, whereas others may not (Labov et al., 1991; Labov, 1994). If a restructuring of phonetic categories takes place on an individual level, and can be modeled as a type of model selection, then we might expect to find empirical variation among members of a speech community is reflected in different model-based clustering predictions about phonetic category structure.

Dutch is a case where a contrast has alternatively claimed to be both neutralized (Lahiri et al., 1987) and incompletely neutralized (Warner et al., 2004), suggesting that this may be a case where considerable individual variation obtains. Compelling evidence for the *in*complete neutralization of this contrast is provided by the studies of Warner et al., who show that the contrast between word-final /t/ and /d/ is not only distinguished by small but statistically significant differences in the realization of acoustic cues such as the duration of the stop burst, but that listeners are able to identify forms such as those in Table 5.1 on the basis of other cues which appear to overlap significantly in production, such as vowel duration and the degree of voicing during closure. Figure 5.6 shows the distribution of four temporal cues to the voicing contrast in final position for words with the long, non-high vowels /a e o/ preceding the final stop: duration of preceding vowel, duration of the stop closure, the length of the voicing period during the closure, and duration of the burst itself.

Warner et al. (2004) compared the realizations of cues to the /t/ $\sim$ /d/ contrast in word-final (neutralizing) position to their realizations in word-medial (non-neutralizing) position. They found that vowel duration was a significant predictor of underlying voicing in both environments, while burst duration was a significant predictor in the supposedly neutralizing environment only in cases where the preceding vowel was long. Neither duration of voicing during the closure nor of the closure itself emerged as significant predictors of underlying voicing in either environment; however, listeners were able to use these cues to discriminate

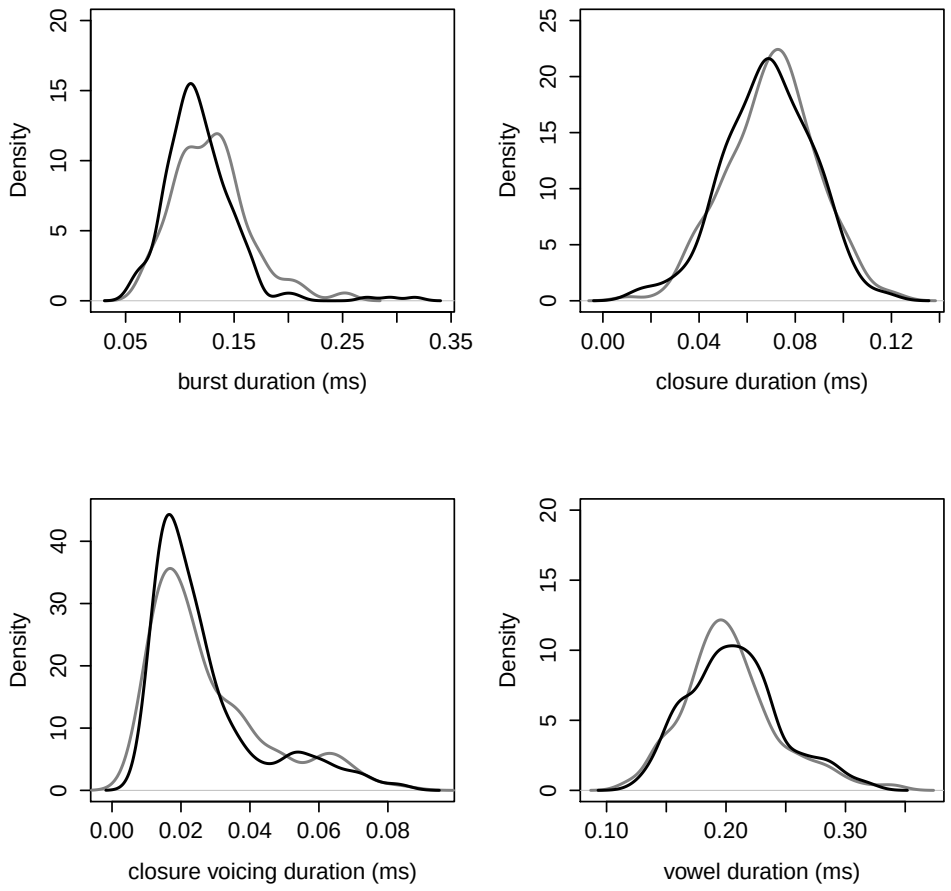

Figure 5.6: Distribution of 4 acoustic cues to Dutch underlying /t/, /d/ in final position for items containing long non-high vowels. Black lines give distribution of underlyingly voiceless stops, gray lines underlyingly voiced stops. Based on data from Warner et al. (2004).

between categories in a two-alternative forced choice perception experiment when all other predictors were held constant. These results suggest that while cues which strongly covary with an underlying phonological contrast will be important for a learner attempting to recover that contrast, other cues which covary less strongly nonetheless play a role in category formation and maintenance.

Warner et al. (2004) also found considerable variation in the individual productions of

these cues, and considered the possibility that an individual speaker's tendency to produce a cue value which covaried with the underlying distinction in production would be related to their ability to use that cue to distinguish the contrast in perception. While the results were not statistically significant, they suggested that listeners who themselves produce larger or more consistent differences in one cue may be more sensitive to the distinction as produced by others. This is consistent with findings that show individual speakers of a given language and even a given speech community vary in their ability to distinguish lexical items based on the degree to which the merger obtains in an individual's own speech (Labov et al., 1991). To the extent that BIC-based model selection makes accurate predictions about human categorization behavior, we should expect to find variation when fitting models to individual speaker data.

5.5.1 The data: Dutch final devoicing

To explore the effects of individual experience on the likelihood of phonetic category restructuring, the experiments described in the previous section were repeated for a subset of the Warner et al. (2004) data.[5] The original data set consists of 1,080 Dutch words, representing two repetitions each by 15 native speakers of 76 Dutch lexical items forming 38 minimal pairs. The production data set contained 300 items with phonologically short and 300 items with phonologically long vowels differing in the underlying voicing of a final obstruent (the NEUTRALIZATION CONTEXT) and 180 items with phonologically short vowels and 300 items with phonologically long vowels differing in the underlying voicing of a medial obstruent (the DISTINCTION CONTEXT). Since the phonologically long vowels /i/ and /u/ are phonetically realized as short (Booij, 1999), only data from non-high long vowels in the NEUTRALIZATION CONTEXT were used in the simulations discussed here; these are shown in Table 5.8.

5. All data used in experiments of §5.5 were collected by Natasha Warner, Allard Jongman, Joan Sereno, and Rachèl Kemps, original discussed and analyzed in Warner et al. (2004). Their willingness to provide these data for analysis here is gratefully acknowledged; however, any errors in the presentation or analysis

orthography	*UR*	*gloss*	*orthography*	*UR*	*gloss*
baat	/bat/	'benefit'	baad	/bad/	'bathe 1st-sg'
boot	/bot/	'boat'	bood	/bod/	'offered 1st-sg'
eet	/et/	'eat sg.'	eed	/ed/	'oath'
meet	/met/	'measure sg.'	meed	/med/	'avoided 1st-sg'
noot	/not/	'nut'	nood	/nod/	'necessity'
smeet	/smet/	'threw sg.'	smeed	/smed/	'forge 1st-sg'
zweet	/zwet/	'sweat'	Zweed	/zwed/	'Swede'

Table 5.8: Experimental items from Warner et al. (2004) used in clustering experiments.

Warner et al. (2004) took acoustic temporal measurements of four cues to underlying final stop voicing: the duration of the release burst (BURST), the duration of the preceding vowel (VDUR), the duration of the closure (CDUR), and the duration of the voiced period of the closure (VGCL). The means, standard deviations, and cue reliability scores ω_d for the subset of the data considered here are shown in Table 5.9.

Category	BURST	VDUR	VGCL	CDUR
voiced	118 (36)	207 (39)	27 (16)	69 (18)
voiceless	130 (34)	208 (40)	28 (17)	71 (19)
ω	0.25	0.37	0.12	0.26

Table 5.9: Parameter values and reliability scores ω for cues to Dutch final stops, non-high neutralization context of Warner et al. (2004) data, all speakers. BURST = burst duration, VDUR = preceding vowel duration, VGCL = duration of voiced period during stop closure, CDUR = duration of closure.

5.5.2 Series 1: Optimality

The first set of simulations was designed to assess the learnability of the Dutch voicing contrast on the basis of the pooled non-high long vowel data from all of the participants in

discussed here should be attributed solely to the present author.

the Warner et al., data set. The procedure was identical to that described in §5.4.1. A series of GMMs were fit to a single set of $N = 500$ 5-dimensional observation vectors generated from a Gaussian mixture with the parameters given in Table 5.9; the BIC scores of models with access to all possible unordered combinations of cue dimensions were then compared. The model-fitting procedure and observation data were identical in all experiments; only the number of cue vector dimensions available to the learner changed.

The results, shown in Table 5.10, indicate that a single-component model is preferred on the BIC metric in all cases, despite marginal improvements in classification accuracy for models with more components.

cue dimension				$K = 1$		$K = 2$		$K = 3$	
d_1	d_2	d_3	d_4	*BIC*	*err*	*BIC*	*err*	*BIC*	*err*
BURST				**1615**	**0.50**	1679	0.48	1664	0.49
VDUR				**1510**	**0.50**	1530	0.49	1521	0.49
VGCL				**2242**	**0.50**	2481	0.47	2497	0.39
CDUR				**2146**	**0.50**	2134	0.47	2123	0.49
BURST	VDUR			**3137**	**0.50**	3217	0.44	3227	0.48
BURST	VGCL			**3857**	**0.50**	4133	0.46	4155	0.40
BURST	CDUR			**3767**	**0.50**	3832	0.49	3822	0.46
VDUR	VGCL			**3784**	**0.50**	4021	0.47	4034	0.49
VDUR	CDUR			**3656**	**0.50**	3672	0.49	3665	0.48
VGCL	CDUR			**4389**	**0.50**	4616	0.47	5069	0.47
BURST	VDUR	VGCL		**5405**	**0.50**	5689	0.45	5719	0.47
BURST	VDUR	CDUR		**5286**	**0.50**	5346	0.47	5345	0.47
BURST	VGCL	CDUR		**6003**	**0.50**	6222	0.47	6297	0.46
VDUR	VGCL	CDUR		**5924**	**0.50**	6146	0.47	6176	0.48
BURST	VDUR	VGCL	CDUR	**7548**	**0.50**	7756	0.48	7846	0.45

Table 5.10: BIC scores and error rates for models in 1–4 dimensions, full Dutch non-low long vowel final neutralization environment. K = number of categories (components); columns show the cue dimensions made available in the observation data. Bold items indicate the optimal solutions. BIC values rounded to nearest integer value. Bayes error=0.40.

5.5.3 *Series 2: Typicality*

The results of the experiments in §5.5.2 appear to run counter to the Warner et al. (2004) results showing that listeners were able to distinguish between the categories with greater-than-chance accuracy. Given that the optimal number of clusters in a model was shown to vary with the particular set of observations to which it was fit for the Korean data, the typicality experiments of §5.4.2 were repeated for the pooled Dutch data to determine if the results of §5.5.2 should be interpreted as indicative of the overall optimality of a single-category solution or perhaps reflect an idiosyncratic statistical property of the particular observation data to which the models were fit. The general model fitting procedure was the same as in §5.4.2; each model parameterization was fit to 1,000 different series of observation vectors drawn from a GMM with parameters again determined from Table 5.9. The BIC-optimal number of model components was recorded for each fit.

The results of the Series 2 typicality experiments confirm the results of §5.5.2: the optimal solution was a single-component model in all cases except for the cue combination BURST + VGCL + CDUR, where $K = 1$ models were selected 90% of the time and $K = 2$ models just 10% (because of the consistency of the results, no tabular representation is provided).

5.5.4 *Series 3–4: Individual optimality and typicality*

The results of the model-fitting experiments in §5.5.2 and §5.5.3 predict complete neutralization of the Dutch voicing contrast in final position, *contra* the results of Warner et al. (2004), who found that listeners were able to distinguish between voicing categories with greater-than-chance accuracy. One possible explanation for this discrepancy is individual differences.[6] As Warner et al. (2004) pointed out, there was a certain amount of variability present in their data among speakers; by pooling the production data for all speakers, this

6. Another possible explanation may be the use of the BIC as a model selection criterion. Future expansions of this work will explicitly compare the results presented here with those based on other approaches to model selection.

individual variation may be lost. In an experiment testing the perceptibility of final voicing, using productions of a subset of their full data set as stimuli, Warner et al. (2004) found that the accuracy with which participants were able to distinguish items differing in underlying voicing varied with the degree to which the categories were distinguished in the stimuli with which they were presented, as might be expected. This is consistent with the hypothesis that phonetic restructuring takes place on an individual level, rather than (or in addition to) at a population level. To explore the range of individual variability in this data, the optimality and typicality experiments were repeated for each of the speakers singled out in Warner et al. (2004) (subjects 3, 5, 6, and 14).

Table 5.11 gives the means, standard deviations, and cue reliability statistics for these speakers; Figure 5.7 gives the individual density plots. The BIC scores and accuracy (error) rates for various combinations of cue dimensions were computed for each of the four subjects. These results are shown in Tables 5.12 – 5.15. Table 5.16 gives an overview of the typicality of solutions for all four individual subjects.

	BURST				VDUR			
	s_3	s_5	s_6	s_{14}	s_3	s_5	s_6	s_{14}
voiced	159 (30)	183 (39)	105 (29)	124 (16)	216 (24)	211 (19)	174 (13)	118 (14)
voiceless	109 (29)	149 (60)	104 (18)	113 (16)	224 (21)	223 (28)	176 (18)	183 (18)
ω	0.22	0.16	0.20	0.25	0.46	0.43	0.44	0.40
	VGCL				CDUR			
	s_3	s_5	s_6	s_{14}	s_3	s_5	s_6	s_{14}
voiced	26 (8)	21 (7)	25 (10)	19 (8)	82 (11)	73 (11)	69 (12)	94 (13)
voiceless	24 (10)	23 (6)	23 (10)	21 (8)	86 (30)	70 (14)	75 (14)	88 (10)
ω	0.13	0.15	0.11	0.08	0.19	0.26	0.25	0.27

Table 5.11: Parameter values and reliability scores ω for cues to Dutch final stops, individual speakers.

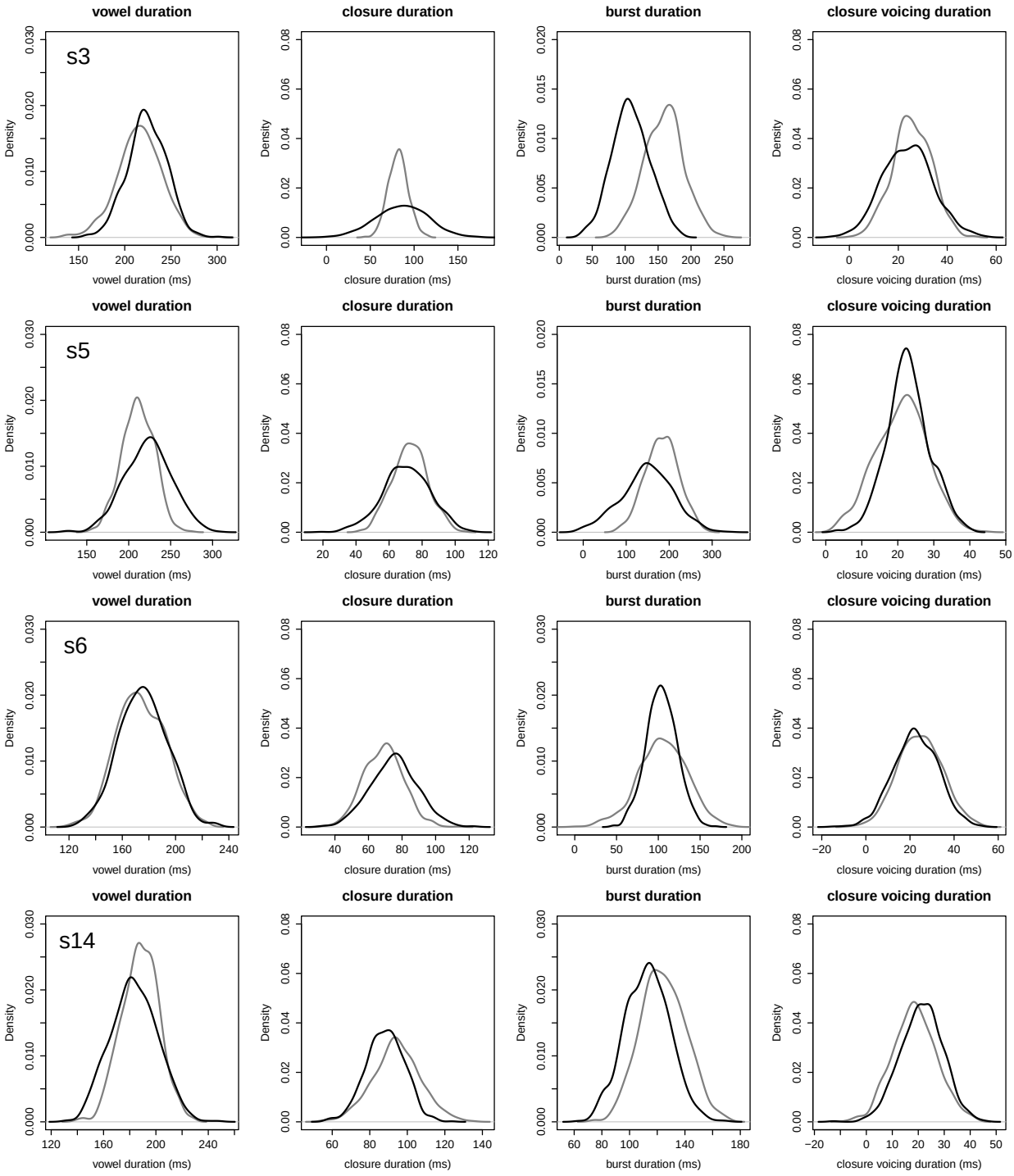

Figure 5.7: Distribution of 4 acoustic cues to underlying /t/, /d/ in final position for items containing long non-low vowels for 4 individual Dutch speakers. Black lines give distribution of underlyingly voiceless stops, gray lines underlyingly voiced stops.

cue dimension				$K = 1$		$K = 2$		$K = 3$	
d_1	d_2	d_3	d_4	*BIC*	*err*	*BIC*	*err*	*BIC*	*err*
BURST				15233	0.50	**15223**	**0.20**	15241	0.47
VDUR				**13680**	**0.50**	13691	0.43	13701	0.47
VGCL				**10922**	**0.50**	10938	0.46	10952	0.44
CDUR				13626	0.50	**13509**	**0.28**	13530	0.35
BURST	VDUR			29806	0.50	**28896**	**0.20**	28918	0.19
BURST	VGCL			26154	0.50	**26126**	**0.24**	26133	0.40
BURST	CDUR			28859	0.50	**28517**	**0.16**	28551	0.16
VDUR	VGCL			**24601**	**0.50**	24620	0.44	24641	0.42
VDUR	CDUR			27299	0.50	**27187**	**0.31**	3665	0.34
VGCL	CDUR			24548	0.50	**24425**	**0.28**	24460	0.41
BURST	VDUR	VGCL		39834	0.50	**39796**	**0.19**	39838	0.41
BURST	VDUR	CDUR		42539	0.50	**42187**	**0.14**	42229	0.36
BURST	VGCL	CDUR		39781	0.50	**39401**	**0.15**	39452	0.18
VDUR	VGCL	CDUR		38228	0.50	**38103**	**0.27**	38151	0.30
BURST	VDUR	VGCL	CDUR	53460	0.50	**53071**	**0.15**	53129	0.24

Table 5.12: BIC scores and error rates for models in 1–4 dimensions, subject s_3. K = number of categories (components); columns show the cue dimensions made available in the observation data. Bold items indicate the optimal solutions. BIC values rounded to nearest integer value. Bayes error = 0.14.

5.5.5 Discussion

The results of the Series 3 and 4 model fitting procedures indicate that both the number of components in the BIC-optimal solutions as well as the typicality of those solutions may vary on a speaker-by-speaker basis. In some cases, the results highlight the potential fragility of the category distinction even for speakers for which it appears to be relatively robust. For instance, consider speaker 3, for whom 2-category solutions are typically BIC-optimal across a range of cue combinations. While the differences in accuracy between 1 and 2-category solutions is typically quite large, as are the differences in BIC, the optimal solutions shown

cue dimension				$K = 1$		$K = 2$		$K = 3$	
d_1	d_2	d_3	d_4	*BIC*	*err*	*BIC*	*err*	*BIC*	*err*
BURST				16196	0.50	**16151**	**0.41**	16164	0.42
VDUR				**13906**	**0.50**	13916	0.38	13926	0.43
VGCL				**9967**	**0.50**	9982	0.44	9983	0.44
CDUR				**11907**	**0.50**	11923	0.44	11932	0.45
BURST	VDUR			30103	0.50	**30051**	**0.29**	30075	0.38
BURST	VGCL			26163	0.50	**26140**	**0.35**	16164	0.37
BURST	CDUR			28104	0.50	**28058**	**0.34**	28085	0.49
VDUR	VGCL			**23873**	**0.50**	23894	0.40	23914	0.49
VDUR	CDUR			**25814**	**0.50**	25818	0.39	25853	0.47
VGCL	CDUR			**21874**	**0.50**	21895	0.45	21916	0.44
BURST	VDUR	VGCL		40069	0.50	**40032**	**0.33**	40055	0.39
BURST	VDUR	CDUR		42010	0.50	**41955**	**0.40**	41968	0.31
BURST	VGCL	CDUR		38070	0.50	**38037**	**0.39**	38049	0.40
VDUR	VGCL	CDUR		**35781**	**0.50**	35807	0.50	35836	0.41
BURST	VDUR	VGCL	CDUR	51977	0.50	**51918**	**0.30**	51974	0.31

Table 5.13: BIC scores and error rates for models in 1–4 dimensions, subject s_5. K = number of categories (components); columns show the cue dimensions made available in the observation data. Bold items indicate the optimal solutions. BIC values rounded to nearest integer value. Bayes error = 0.26.

in Table 5.12 for combinations such as BURST + VDUR, where the contrast is particularly well-separated along the dimension BURST, are relatively untypical. While this once again underscores both the role of individual experience as well as the importance of considering higher-dimensional separability in predicting phonetic category restructuring, it also highlights the stochastic nature of the restructuring process.

In addition, note that while a category contrast may be predicted to be neutralized (i.e., the BIC-optimal solution is $K = 1$) when a single cue dimension is considered, that prediction is often reversed as more cue dimensions are considered. However, this is not always the case: for some speakers, while training on two-dimensional observations may

cue dimension				$K = 1$		$K = 2$		$K = 3$	
d_1	d_2	d_3	d_4	*BIC*	*err*	*BIC*	*err*	*BIC*	*err*
BURST				13864	0.50	**13862**	**0.39**	13864	0.43
VDUR				12977	0.50	**12987**	**0.48**	13003	0.48
VGCL				**11212**	**0.50**	11228	0.46	11243	0.47
CDUR				**12043**	**0.50**	12058	0.39	12067	0.45
BURST	VDUR			**26842**	**0.50**	26865	0.48	26867	0.44
BURST	VGCL			25706	0.50	**25089**	**0.47**	25106	0.49
BURST	CDUR			25907	0.50	**25904**	**0.41**	25921	0.46
VDUR	VGCL			**24188**	**0.50**	24204	0.47	24226	0.41
VDUR	CDUR			**25020**	**0.50**	25034	0.48	25051	0.40
VGCL	CDUR			**23254**	**0.50**	23270	0.40	23289	0.44
BURST	VDUR	VGCL		**38053**	**0.50**	38074	0.50	38099	0.44
BURST	VDUR	CDUR		**38884**	**0.50**	38907	0.48	38927	0.46
BURST	VGCL	CDUR		37119	0.50	**37116**	**0.38**	37158	0.40
VDUR	VGCL	CDUR		**36231**	**0.50**	36252	0.49	36281	0.48
BURST	VDUR	VGCL	CDUR	**50096**	**0.50**	50117	0.46	50145	0.48

Table 5.14: BIC scores and error rates for models in 1–4 dimensions, subject s_6. K = number of categories (components); columns show the cue dimensions made available in the observation data. Bold items indicate the optimal solutions. BIC values rounded to nearest integer value. Bayes error = 0.35.

typically result in BIC-optimal solutions where $K = 2$, adding a third dimension can cause a reversal, resulting in a preponderance of $K = 1$ solutions (compare e.g. speaker 5 BURST + CDUR with BURST + VGCL + CDUR, Table 5.16). More generally (and consistent with the Korean and pooled Dutch data considered above), access to a cue dimension that, when considered on its own, predicts a K-category solution to be optimal does not necessarily imply that a K-category solution will be optimal when considered simultaneously with other dimensions.

cue dimension				$K = 1$		$K = 2$		$K = 3$	
d_1	d_2	d_3	d_4	*BIC*	*err*	*BIC*	*err*	*BIC*	*err*
BURST				**12811**	**0.50**	12825	0.37	12838	0.41
VDUR				**12642**	**0.50**	12655	0.42	12669	0.44
VGCL				**10649**	**0.50**	10664	0.42	10679	0.46
CDUR				**11708**	**0.50**	11724	0.39	11729	0.46
BURST	VDUR			**25447**	**0.50**	25455	0.39	25473	0.45
BURST	VGCL			**23460**	**0.50**	23484	0.43	23505	0.42
BURST	CDUR			**24519**	**0.50**	24536	0.39	24548	0.41
VDUR	VGCL			**23288**	**0.50**	23305	0.42	23326	0.44
VDUR	CDUR			**24350**	**0.50**	24365	0.41	24386	0.47
VGCL	CDUR			**22356**	**0.50**	22374	0.39	22380	0.45
BURST	VDUR	VGCL		**36100**	**0.50**	36102	0.42	36130	0.42
BURST	VDUR	CDUR		**37161**	**0.50**	37169	0.40	37192	0.47
BURST	VGCL	CDUR		**35168**	**0.50**	35180	0.39	35207	0.40
VDUR	VGCL	CDUR		**34999**	**0.50**	35009	0.44	35036	0.44
BURST	VDUR	VGCL	CDUR	**47810**	**0.50**	47811	0.40	47847	0.439

Table 5.15: BIC scores and error rates for models in 1–4 dimensions, subject s_{14}. K = number of categories (components); columns show the cue dimensions made available in the observation data. Bold items indicate the optimal solutions. BIC values rounded to nearest integer value. Bayes error = 0.30.

5.6 General discussion

The model-based clustering experiments conducted in this chapter suggest that the likelihood of phonetic category restructuring may not be solely a function of accuracy (model fit), but may represent an adaptive trade-off between accuracy and model complexity. To the extent that BIC-based model selection may be assumed as an adequate model of human clustering behavior, this means that neutralization of a contrast is not a foregone conclusion simply because accuracy is compromised.

The results of the optimality experiments fitting to both the Dutch and Korean data also

cue dimension				s_3 *number of components*			s_5 *number of components*		
d_1	d_2	d_3	d_4	$K=1$	$K=2$	$K=3$	$K=1$	$K=2$	$K=3$
BURST				**0.66**	0.34		0.19	**0.81**	
VDUR				**0.98**	0.02		**0.75**	0.25	
VGCL				**1**			**0.99**	0.01	
CDUR					**1**		**0.99**	0.01	
BURST	VDUR			**0.81**	0.18	0.01	0.04	**0.96**	
BURST	VGCL			**0.64**	0.32	0.04	0.39	**0.61**	
BURST	CDUR				**0.98**	0.02	0.35	**0.65**	
VDUR	VGCL			**1**			**0.82**	0.18	
VDUR	CDUR				**0.99**	0.01	**0.71**	0.29	
VGCL	CDUR				**0.98**	0.02	**1**		
BURST	VDUR	VGCL		**0.55**	0.44	0.01	0.20	**0.80**	
BURST	VDUR	CDUR			**0.98**	0.02	0.03	**0.97**	
BURST	VGCL	CDUR			**1**		**0.58**	0.42	
VDUR	VGCL	CDUR			**0.98**	0.02	**0.91**	0.09	
BURST	VDUR	VGCL	CDUR		**0.99**	0.01	0.19	**0.80**	0.01

				s_6				s_{14}	
BURST				**0.88**	0.12		**1**		
VDUR				**0.99**	0.01		**0.99**	0.01	
VGCL				**1**			**1**		
CDUR					**0.99**	0.01	**0.92**	0.08	
BURST	VDUR			**0.98**	0.02		**1**		
BURST	VGCL			**0.98**	0.02		**1**		
BURST	CDUR				**0.86**	0.14	**0.97**	0.03	
VDUR	VGCL			**1**			**0.98**	0.02	
VDUR	CDUR				**1**		**0.93**	0.07	
VGCL	CDUR				**0.98**	0.02	**0.98**	0.02	
BURST	VDUR	VGCL		**1**			**0.95**	0.05	
BURST	VDUR	CDUR			**0.95**	0.05	**0.91**	0.09	
BURST	VGCL	CDUR			**0.93**	0.06	0.01	**0.97**	0.03
VDUR	VGCL	CDUR			**1**		**0.99**	0.01	
BURST	VDUR	VGCL	CDUR		**0.97**	0.03	**0.88**	0.12	

Table 5.16: Typicality of BIC-optimal category solutions for individual Dutch speakers, non-low long vowel neutralization environment data in terms of percentage of 1,000 fits. Most-typical (≥ 0.50) solution percentages given in bold.

illustrate the point that statistical separability of a contrast cannot necessarily be inferred from separability along individual acoustic-phonetic dimensions. While this is unsurprising from a statistical standpoint, it is important to highlight it in the context of phonetic classification and categorization, because it underscores the need to consider not just sensitivity to individual acoustic dimensions, but also to determining which cue dimensions are relevant and how those dimensions are weighted by individual listeners. As shown by the typicality experiments, however, small variations in the distribution of the observation data to which the statistical learner is exposed can have a considerable impact on the number of the components in the optimal solution. In the case of the pooled Dutch data considered in §5.5.2–5.5.3, for instance, while $K = 1$ solutions were generally preferred for observation data which included only BURST and VDUR information, $K = 2$ solutions were more likely to be optimal for other types of 2-dimensional observation data, such as that containing only BURST and VGCL information. When BURST information (the dimension which covaried most robustly with the underlying voicing specification) was unavailable, a model with a single component was nearly always optimal. As the number of cue dimensions made available to the learners is increased, however, interpretation of the results becomes less straightforward. The factors affecting the relative typicality of selecting a BIC-optimal model with one or two components for 3- and 4-dimensional data are not immediately obvious; the relative typicality of solutions did not appear to vary with the presence or absence of any single cue dimension.

The typicality results thus suggest that while access to additional cues *may* assist learners in recovering a covert contrast, there is no guarantee that this will be the case: the higher accuracy afforded models with access to more cue dimensions is only justified if the increase in model complexity is not too great, and the outcome of model selection is heavily influenced by distributional parameters as well as fit/complexity considerations. From the standpoint of human speech perception, this raises the question of if the distributions of cues which overlap

in isolation could nonetheless assist a listener in the classification and clustering of phonetic categories. Furthermore, it suggests the possibility that some experimental findings showing that human listeners cannot discriminate between supposedly neutralized categories at above chance levels may be misleading, in that their design may not allow for a positive outcome. If categories are only recoverable when learners have access to a wide range of varying acoustic cues, then failure of participants to discriminate categories in a traditional two-alternative forced choice paradigm, where one acoustic dimension is varied while all others are held constant, cannot be taken as evidence that the categories are in fact indiscriminable, or even that the acoustic dimension tested plays no role in category discrimination, since a given cue dimension may only prove informative in discriminating a contrast for some particular value of all other relevant cues. Future laboratory exploration of the present results may need to consider alternative experimental approaches to more accurately assess the integration of cues in human categorization and category learning (Plauché, 2001; Pothos and Chater, 2002; Pothos and Close, 2008; Pothos and Bailey, 2009; Goudbeek et al., 2008).

5.6.1 The role of individual variation

Individual speakers of a given language and even a given speech community are known to vary in their productions of the same contrast. In some cases, these are differences in production strategies that produce the same acoustic output, but variation in acoustic output is also frequently observed (see e.g. Newman, Clouse, and Burnham (2001) for differences in the distribution of spectral center of gravity or Allen, Miller, and DeSteno (2003) for differences in the distribution of VOT).

This type of individual variation if clearly present in the production data of Warner et al. (2004). The results of the computational model-fitting experiments using the production data of the individual Dutch subjects highlight the importance of considering the impact of differences in individual speaker experience on phonetic category perception and restruc-

turing. The results presented here are consistent with the findings that some members of a speech community can show covert contrast/near mergers in production, perception, or both, while others neither produce nor perceive such contrasts (Labov et al., 1991). If perceptual separability can be predicted from production in the manner suggested in this dissertation, and if BIC-based model selection is an accurate model of human behavior, then this predicts exactly this type of variability in a population.

There is some experimental evidence that individual-level differences in production may impact the perceptual dimensions attended to by listeners. In assessing the perceptual weight of release bursts and formant transitions taken from syllable-initial voiced stops of two American English speakers, Dorman, Studdert-Kennedy, and Raphael (1977) found significant differences in the perceptual weight accorded the burst between speakers. While the authors interpreted their results in terms of the 'functional invariance' debate, their results also suggests that the set of perceptual cues potentially relevant for discrimination of a contrast may not be attended to, or accorded equal perceptual weight, by all members of a speech community. A similar conclusion may be drawn based on the work of Chandrasekaran et al. (2010), who have shown that listeners' ability to learn lexical tones was a function not only of training to attend to the relevant cues (pitch height and pitch direction) but also to individuals' pre-training ability to attend to those cues. These studies suggest that the RESTRUCTURING PROBLEM is in some sense an individual-level issue, and understanding when phonetic category contrasts are more or less likely to collapse involves a detailed understanding of the perceptual weight different individuals afford to the relevant acoustic dimensions. In terms of the model selection procedure described in this chapter, this means that some individuals, exposed to the same observation data, may attend to different aspects (columns) of that data, and that this in and of itself can lead to different predictions about the overall category structures they may posit.

5.6.2 The restructuring problem

The stated goal of this chapter was to address the RESTRUCTURING PROBLEM as laid out in Chapter 1: under what conditions will loss of contrast precision induce (trans)phonologization, and in what cases will it lead to neutralization? The Korean simulation results of Chapter 4 suggested that the answer would be mainly distributional: if two categories came to overlap considerably along multiple of the acoustic-phonetic dimensions relevant for their accurate perception, this would result in neutralization of the contrast. To test this hypothesis, the internal states of agents where neutralization of categories seemed *a priori* likely were submitted to model-based clustering using the BIC regularization criterion.

At first glance, the results presented here do not appear to support the hypothesis: the BIC-optimal model fit to the output of the enhancement-only Korean simulations had two categories, not one. However, further computational simulations revealed considerable variation in the number of BIC-optimal categories as the number and combination of cue dimensions was manipulated, suggesting that individual variation may play an important role, a conclusion further supported by the results of model-based clustering using the individual Dutch data of Warner et al. (2004).

The results of model-based clustering using both Korean and Dutch data suggest that the RESTRUCTURING PROBLEM actually consists of at least two distinct subproblems: the problem of predicting category restructuring for a given individual in a given situation with given experience; and the problem of determining to what degree adaptive enhancement will ameliorate the effects of phonetic biases of the type discussed in Chapter 4. While the Dutch experiments in particular underscore the important role of individual variation, thus speaking to the first subproblem, they do not directly address the second. Put slightly differently, we might ask: why has Korean transphonologized in the face of contrast elimination, while Dutch has allowed a (near) merger?

One line of explanation lies in the idea that the word-final voicing contrast in Dutch

bears less functional load than the word-initial voicing contrast in Korean, and therefore that positional neutralization in the Dutch case has less impact on the system of phonological contrasts than it would in Korean (Kingston, 2007). In terms of the modeling approach taken here, these differences would be accounted for by differences in the β parameter used to indicate functional load. At present, however, this is a free parameter of the model, and moreover, it is unclear how functional load should be measured. A major challenge for future research will be to determine how to empirically ground the β parameter in much the same way that the ω and ϵ parameters have been empirically grounded in distributional cue statistics.

It is also important to point out that although factors such as functional load are in some sense outside the purview of the present model, they are not incompatible with it. In other words, we must recognize that are other factors relevant to this aspect of the RESTRUCTURING PROBLEM, such as population dynamics and lexical idiosyncrasies, that will impact the degree to which enhancement compensates for bias. These types of factors are not addressed by the current model, which is focused on the perceptual-distributional aspects of category formation. This is not to imply in any way that extraperceptual factors are not important, but simply to note that this indeterminacy on the part of the present model should not be regarded as a fatal flaw: it provides only a partial answer to a larger and more complex question.

5.7 Conclusion

This chapter has proposed a model-based clustering approach to the RESTRUCTURING PROBLEM, whereby the structure of the phonetic category space is determined by a trade-off between model fit and model complexity. The results of a number of experiments fitting models to empirical data were used to illustrate that the likelihood of phonetic category restructuring, as modeled by number of categories posited by an unsupervised statistical

learner, is dependent on the type and number of cues provided in the observation data. When given access only to cues which individually covaried reliably in production with an underlying phonological contrast, statistical learners were in some cases unable to learn the underlying distribution; in some (but not all) cases, access to additional cue dimensions facilitated category learning. This suggests that a given individual's phonetic experience may lead them to predict different category-level structures, leading in turn to variation within a speech community. Taking a view of category induction as a statistical learning problem is a useful means of parameterizing the factors involved in determining when the number of phonetic categories may change for a given learner, and can help to better understand how those factors interact.

CHAPTER 6

SUMMARY AND CONCLUSIONS

6.1 Summary

This dissertation has considered three questions raised by the phonologization model of sound change, termed the SELECTION, TRADING, and RESTRUCTURING PROBLEMS. Chapter 1 identified and described these problems. The SELECTION PROBLEM was defined as determining which of a number of subphonemic cues is likely to be targeted in phonologization; the closely related TRADING PROBLEM involves determining why phonologization of a redundant cue is invariably accompanied by dephonologization of a primary cue. Finally, the RESTRUCTURING PROBLEM was concerned with changes to higher-level category structures as a result of subphonemic reorganization. It was argued that these questions can be answered by taking an approach to sound change in which the enhancement of phonetic cues and the induction of phonetic categories are both driven by broadly adaptive considerations of robustness (accuracy) and complexity (efficiency).

Chapters 2 and 3 then laid out the details of a formal model in which these adaptive hypotheses could be implemented, in order to gain a better sense of the range of empirical phenomena they would predict. After considering aspects of the speech signal known to be important for auditory perception and classification, Chapter 2 described the framework of FINITE MIXTURE MODELS, illustrating their application to the production and perception of phonetic categories. Chapter 3 detailed an agent-based iterated learning environment in which to model the interaction of multiple speaker/listeners, using the mixture models described in the previous chapter. Chapter 4 employed this simulation framework to implement the PROBABILISTIC ENHANCEMENT HYPOTHESIS and to determine its potential role in explaining the SELECTION and TRADING PROBLEMS in a particular case of phonologization, the transphonologization of F0 in Seoul Korean. In simulations where probabilistic enhance-

ment was implemented in the absence of any threat to contrast precision, the results did not resemble the empirical findings, nor did simulations in which contrast precision was reduced in the absence of a probabilistic enhancement strategy. However, when both factors were simulated, the results closely resembled the empirical findings.

Chapter 5 addressed the RESTRUCTURING PROBLEM by way of analogy with the problem of regularization in model selection. In truly unsupervised learning scenarios, the number of category labels is not given in advance, but must be inferred based on distributional properties of the observed data, subject to some type of criterion. Using one particular criterion, Chapter 5 demonstrated that contrasts may not be as threatened by distributional overlap along multiple cue dimensions as may be supposed, on account of a contrast remaining separable in a higher-dimensional space, and that this may account for certain cases of INCOMPLETE NEUTRALIZATION. To further explore this possibility, mixture models were fit to production data from a putative instance of incomplete neutralization in Dutch. Considerable diversity in the typicality of the number of components in optimal models across individuals was discovered, consistent with sociolinguistic findings, highlighting a crucial role for individual variation in addressing the RESTRUCTURING PROBLEM.

Overall, the results demonstrate that adaptive enhancement can account for both cue selection as well as the appearance of cue trading in phonologization, and illustrate how changes at the category level may be modeled using the same representational framework as changes at the subphonemic level. This framework was used to model both the transphonologization of F0 in Seoul Korean and the incomplete neutralization of voicing in Dutch word-final obstruents. While by no means proving that these are the definitive means by which these sound change took place, the results demonstrate that it is possible to account for some instances sound change via gradual, adaptive processes of subphonemic reorganization, in which any and all phonetic dimensions are in principle accessible to enhancement, and in which both communicative accuracy and efficiency play a role.

6.2 Outstanding questions and future directions

6.2.1 Individual variation and population dynamics

The focus of this dissertation has been on properties of the individual speaker, and how those properties change over time in response to input and subject to certain perceptual and communicative constraints. In the general case, however, speakers exist and interact within the context of a larger speech community, and the dynamics of population interaction and evolution will necessarily influence the nature of the input to any given individual. While the results of Chapter 5 highlight the potential impact of individual variation on phonetic category restructuring, the magnitude of such impact will be mitigated and influenced by the range of individuals that a given agent comes into contact with. Thus the fate of a contrast, or the informativeness of a cue, must be considered both at the individual level, but also at the population level, and it must be recognized that differences in population dynamics may have an impact on the typicality of the results present here (Niyogi, 2006; Niyogi and Berwick, 2009). From the present perspective, it is interesting to consider the possibility that apparent stability in the acoustic realization of phonetic categories may arise at the population level rather than at the level of the individual.

6.2.2 Induction of acoustic-phonetic cues

Although this dissertation briefly considers the problem of inducing phonetic category labels in an unsupervised fashion, the relevant acoustic-phonetic cue dimensions are still assumed to be available *a priori*. It is interesting to consider for a moment the problem of CUE INDUCTION, that is, learning to identify regions of the phonetic record which are likely to act as perceptual cues to high-level contrastive units.

Much that we currently know about perceptual cues comes from the work conducted by Pierre Delattre, Franklin Cooper, Alvin Liberman and their colleagues at Haskins Labs

beginning in the 1950s (see Liberman and Cooper, 1972 for a summary and references). These researchers identified acoustic-phonetic dimensions essentially by manipulating parameters of an early speech synthesizer (Cooper, 1953) and determining if the results would elicit a perceptual response. On this view, a cue is 'a portion of the signal that can be isolated visually, that can be manipulated independently in a speech synthesizer constructed for that purpose, and that can be shown to have some perceptual effect' (Repp, 1982). From the standpoint of the unsupervised identification of perceptually salient units, this definition is rather problematic, in that it presupposes perceptual salience. It would be interesting to try and identify such acoustic parameters on the basis of more general principles, at which point the accuracy of such a system could be assessed by comparison to the formidable baseline assembled by the Haskins team.

One line of research that takes steps in this direction is the work of Ken Stevens and colleagues on landmark detection (Stevens and Blumstein, 1981; Stevens, 1985, 1995; Liu, 1995; Halberstadt and Glass, 1997; Halberstadt, 1998; Juneja, 2004; Hasegawa-Johnson et al., 2005). However, a landmark-based speech recognizer is built on prior domain knowledge of what types of acoustic events will be relevant for the perception and identification of certain classes of sounds; the challenge is to determine how to induce this set of events from properties of the acoustic signal. Some of the recent work of Jont Allen and colleagues addresses itself to this latter problem (Allen and Li, 2009; Allen et al.,2009; Li et al., 2010).

In addition, while the present work proposes that the set of category labels may undergo restructuring in response to changes to the distributions of the individual cue dimensions, Chapter 5 also demonstrated how restructuring may be impacted by changes to the number and combination of distributions. In future work, it will be important to consider the factors that may lead to the restructuring of the cue space itself, in terms of selective attention to different dimensions, and the impact this may have on the space of category labels (Francis and Nusbaum, 2002).

6.2.3 Stage transitions and symbolic representation

The finite mixture representations advocated in the present work, while not at all uncommon in many areas of speech science, are not yet commonplace in the post-generative linguistics literature. From the standpoint of sound change, mixture model representations allow for new types of explanations such as those offered here, which make crucial reference to variation along multiple acoustic-phonetic dimensions, and directly encode the idea that certain dimensions may contribute more to the maintenance of a contrast than others.

However, this representation retains a rather traditional view of the phonetics-phonology interface, in that a set of cues (features) is associated with a particular symbolic unit in an autosegmental sense (Goldsmith, 1976). This brings up the issue of determining whether the gradient, variable subproperties of the symbolic unit (category label) should be reflected in the discrete properties of the label itself.

Stage I	*Stage II*	*Stage III*
pá [⁀—]	pá [⁀—]	pá [⁀—]
bá [⁀⁀]	bă [⁀⌒]	pă [⁀⌒]

Table 6.1: (Trans)phonologization and phonemicization (after Hyman, 1976). Sparklines show the time course of F0 production for the vowel following the initial consonant.

To see how this is problematic, consider again Hyman's three-stage conception of phonologization, reproduced here in Table 6.1. At issue is the difference between Stage II and Stage III; in particular, the choice of the representation [pă] versus [bă]. In Chapter 1, Stage II was described as a point at which 'the pitch differences may be regarded as allophonic, conditioned by the initial consonant, but the stage has been set whereby a reanalysis may occur following the loss of the voicing contrast in the initials' (2). But how are we to determine when this reanalysis has occurred? When, precisely, does the voicing contrast in the initials become 'lost'? If the choice of symbolic representation is arbitrary, then the difference between Stage II and Stage III is not really captured by using a different symbol for the initial

segment; but if the difference is not arbitrary, then exactly what subsymbolic properties are embodied in those representations, and how are they relevant at the phonological level?

It is not immediately clear how to resolve this problem. One possible response is suggested by the gestural representations of Articulatory Phonology (Browman and Goldstein, 1986, 1989; Gafos, 2002), which provide a natural way of expressing non-segmental, continuous, and gradient aspects of phonetic realization. However, as noted by Ladd (to appear), it is not clear if these aspects need to be embedded in phonological abstractions, or how these representations capture many of the cross-linguistic symmetry in phonological patterning that has been central to much of post-Structuralist theoretical phonology; in addition, there are outstanding questions about how this type of gestural representation might be reconciled with many of the perceptual confusion results (Plauché et al., 1997; Plauché, 2001; Guion, 1998). A major challenge for the modern phonological enterprise will be to reconcile what is known about the continuous and variable nature of the phonetic signal with what is known about the behavior of symbolic processes in a representationally consistent fashion.

6.2.4 Sound change in the laboratory

The strategy of this dissertation has been to explore the predictions of certain hypotheses through concrete mathematical formalization and computational simulation. The results form a sufficiency proof in that they demonstrate that such a path to these types of sound changes is in principle possible. Additional supporting evidence may be sought in laboratory experiments with human subjects. Here, I briefly describe some possible experimntal extensions of the present work.

It is well known that listeners are able both to redistribute attention to cues both actively (Francis and Nusbaum, 2002) as well as passively (Kraljic and Samuel, 2006). On the adaptive approach to phonetic change taken in this dissertation, perceptual reorganization is predicted to take place proportional to the distributions of cues in the input. One additional

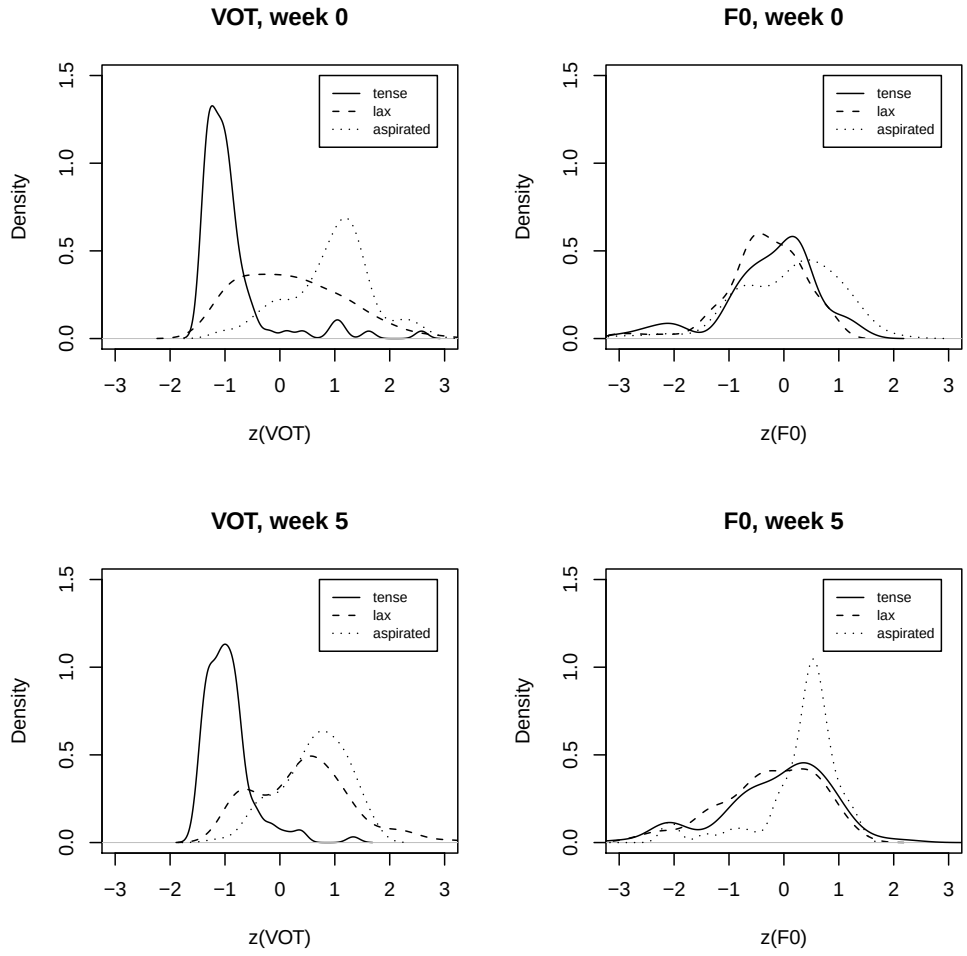

Figure 6.1: Distributions of VOT and F_0 for nonnative Korean learners, prior to receiving instruction (top row) and after 5 weeks of instruction (bottom row). From Kirby and Yu (in prep.).

assumption made in this dissertation, however, is that this reorganization will be reflected both in perception as well as in production. This hypothesis has been investigated by Kirby and Yu (in prep.), in which 17 native speakers of American English with no prior knowledge of Korean performed perception and production tasks at regular interval during their participation in a 10-week intensive Korean language course. Each week, the participants were recorded producing items differing in initial stop place and manner. While participants

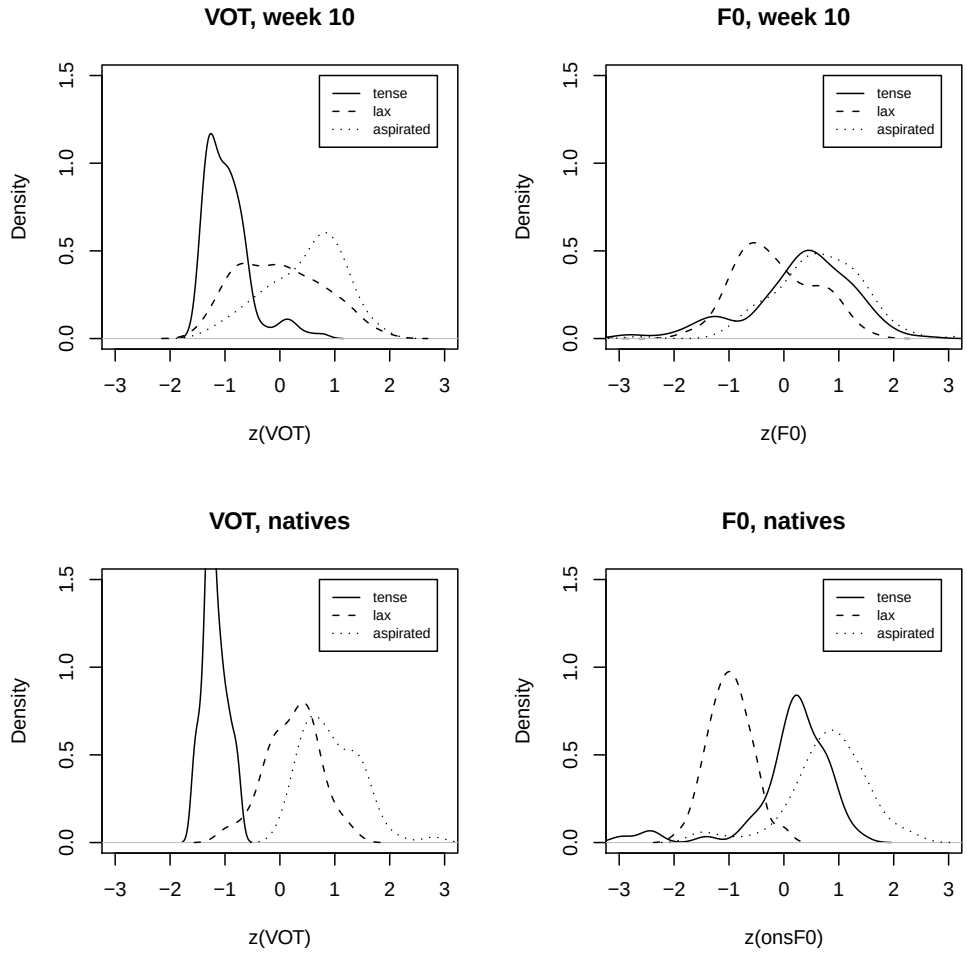

Figure 6.2: Distributions of VOT and F_0 for nonnative Korean learners after 5 weeks of instruction (top row) compared to native Korean controls (bottom row). From Kirby and Yu (in prep.).

initially failed to distinguish the three-way Korean stop contrast in production in a manner consistent with the input they were receiving, over time their cue distributions came to approximate those of the input to a greater and greater degree. In particular, separation between lax and aspirated/tense categories increased along the F0 dimension, while separation between lax and aspirated categories decreased along the VOT dimension, as illustrated in Figures 6.1 and 6.2. This result demonstrates that listeners are able to reorganize their cue

space in a manner consistent with the input even without the presence of explicit feedback, and that this reorganization is reflected in their own productions of those categories.

In addition, the PROBABILISTIC ENHANCEMENT HYPOTHESIS seems particularly well-suited for laboratory investigation. Here, it would need to be demonstrated not simply that participants produce categories in a manner consistent with the subphonemic structure of the input, but that they enhance aspects of the signal proportional to their reliability. The challenge will be to demonstrate that those portions of the signal are enhanced based on their informativeness, as opposed to being able to attribute enhancement to some other factor.

6.3 Conclusions

This dissertation has proposed a framework for modeling and exploring sound change using a multidimensional representation which retains rather than discards within-category variation in order to address three problems of SELECTION, TRADING, and RESTRUCTURING. The answer to the problem of cue SELECTION is that cues are targeted and enhanced based on their reliability; the appearance of cue TRADING emerges as a byproduct of finite and relative attentional resources; and category RESTRUCTURING takes place on the basis of changes to distributional subphonemic properties. By providing several examples of how this type of representation can be applied, it is hoped that this work may serve as a template for others to improve upon in addressing trenchant problems in linguistic theory and sound change.

REFERENCES

Abramson, A. S. and L. Lisker (1972). Voice timing in Korean stops. In *Proceedings of the Seventh International Congress of Phonetic Sciences*, The Hague, pp. 439–446. Mouton.

Ahn, H. (1999). *Post-release phonatory processes in Korean and English: Acoustic correlates and implications for Korean phonology*. Ph. D. thesis, University of Texas at Austin.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control 19*(6), 716–723.

Allen, J. B. and F. Li (2009). Speech perception and cochlear signal processing. *IEEE Signal Processing Magazine 29*, 117–123.

Allen, J. B., M. Régnier, S. Phatak, and F. Li (2009). Nonlinear cochlear signal processing and phoneme perception. In N. P. Cooper and D. T. Kemp (Eds.), *Proceedings of the 10th Mechanics of Hearing Workshop*, pp. 93–105. World Scientific Publishing Co.

Allen, J. S., J. L. Miller, and D. DeSteno (2003). Individual talker differences in voice-onset-time. *Journal of the Acoustical Society of America 113*(1), 544–552.

Anttila, R. (1989). *An Introduction to Historical and Comparative Linguistics* (2nd ed.), Volume 4 of *Current Issues in Linguistic Theory*. Amsterdam: John Benjamins.

Ashby, F. G. (1992). Multidimensional models of categorization. In F. G. Ashby (Ed.), *Multidimensional Models of Perception and Cognition*, pp. 449–483. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Ashby, F. G. and L. A. Alfonso-Reese (1995). Categorization as probability density estimation. *Journal of Mathematical Psychology 39*, 216–233.

Ashby, F. G. and W. T. Maddox (1993). Relations between prototype, exemplar, and decision bound models of categorization. *Journal of Mathematical Psychology 37*, 372–400.

Ashby, F. G. and N. A. Perrin (1988). Toward a unified theory of similarity and recognition. *Psychological Review 95*(1), 124–150.

Ashby, F. G., S. Queller, and P. Berretty (1999). On the dominance of unidimensional rules in unsupervised categorization. *Perception & Psychophysics 61*, 1178–1199.

Ashby, F. G. and E. M. Waldron (1999). On the nature of implicit categorization. *Psychonomic Bulletin & Review 6*(3), 363–378.

Aylett, M. P. and A. Turk (2004). The Smooth Signal Redundancy Hypothesis: A functional explanation for relationships between redundancy, prosodic prominence and duration in spontaneous speech. *Language and Speech 47*(1), 31–56.

Baker, A. (2008). Addressing the actuation problem with quantitative models of sound change. In *Proceedings of the 31st Annual Penn Linguistics Colloquium*, Volume 14(1), pp. 29–41.

Baudouin de Courtenay, J. (1895). *Versuch einer Theorie phonetischer Alternationen: Ein Kapitel aus der Psychophonetik*. Strassburg: K. J. Trübner.

Baxter, W. H. (1992). *A Handbook of Old Chinese Phonology*. Berlin: Mouton de Gruyter.

Beddor, P. S. (2009). A coarticulatory path to sound change. *Language 85*(4), 785–821.

Beddor, P. S., J. D. Harnsberger, and S. Lindemann (2002). Language-specific patterns of vowel-to-vowel coarticulation: acoustic structures and their perceptual correlates. *Journal of Phonetics 30*(4), 591–627.

Beddor, P. S., R. A. Krakow, and L. M. Goldstein (1986). Perceptual constraints and phonological change: a study of nasal vowel height. *Phonology Yearbook 3*, 197–217.

Best, C. T., G. W. McRoberts, and N. M. Sithole (1988). Examination of perceptual reorganization for nonnative speech contrasts: Zulu click discrimination by English-speaking adults and infants. *Journal of Experimental Psychology: Human Perception and Performance 14*(3), 345–360.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. New York: Springer Verlag.

Blevins, J. (2004). *Evolutionary Phonology*. Cambridge: Cambridge University Press.

Blevins, J. (2006). A theoretical synopsis of Evolutionary Phonology. *Theoretical Linguistics 32*(2), 117–166.

Blevins, J. and A. Garrett (1998). The origins of consonant-vowel metathesis. *Language 74*(3), 508–556.

Blevins, J. and A. Wedel (2009). An evolutionary approach to lexical competition. *Diachronica 26*(2), 143–183.

Bloomfield, L. (1933). *Language*. New York: Henry Holt.

de Boer, B. (2000). Self organization in vowel systems. *Journal of Phonetics 28*(4), 441–465.

de Boer, B. (2001). *The Origins of Vowel Systems*. Oxford: Oxford University Press.

de Boer, B. and P. Kuhl (2003). Investigating the role of infant-directed speech with a computer model. *Acoustics Research Letters On-line 4*(4), 129–134.

de Boer, B. and W. Zuidema (2009). The evolution of combinatorial phonology. *Journal of Phonetics 37*(2), 123–144.

Boersma, P. (1998). *Functional Phonology*. Ph. D. thesis, University of Amsterdam.

Booij, G. (1999). *The Phonology of Dutch*. Oxford: Clarendon Press.

Bosch, L. and N. Sebastián-Gallés (2003). Simultaneous bilingualism and the perception of a language-specific vowel contrast in the first year of life. *Language and Speech 46*(2–3), 217–243.

Bradlow, A. R. (2002). Confluent talker- and listener-related forces in clear speech production. In C. Gussenhoven and N. Warner (Eds.), *Laboratory Phonology 7*, pp. 241–272. Berlin: Mouton de Gruyter.

Bradlow, A. R. and T. Bent (2002). The clear speech effect for non-native listeners. *Journal of the Acoustical Society of America 112*(1), 272–284.

Browman, C. P. and L. Goldstein (1986). Towards an articulatory phonology. *Phonology Yearbook 3*, 219–252.

Browman, C. P. and L. Goldstein (1989). Articulatory gestures as phonological units. *Phonology 6*, 201–251.

Bybee, J. (2001). *Phonology and Language Use*. Cambridge: Cambridge University Press.

Campbell, L. (1996). On sound change and challenges to regularity. In *The Comparative Method Reviewed: Regularity and Irregularity in Language Change*, pp. 72–89. Oxford: Oxford University Press.

Campbell, L. (1998). *Historical Linguistics: An Introduction*. Cambridge, MA: The MIT Press.

Chandrasekaran, B., P. D. Sampath, and P. C. M. Wong (2010). Individual variability in cue-weighting and lexical tone learning. *Journal of the Acoustical Society of America 128*(1), 456–465.

Chang, S.-E. (2007). *The phonetics and phonology of South Kyungsang Korean tones*. Ph. D. thesis, University of Texas at Austin.

Cho, T. (1996). Vowel correlates to consonant phonation: An acoustic-perceptual study of Korean obstruents. Master's thesis, University of Texas at Arlington.

Cho, T., S.-A. Jun, and P. Ladefoged (2002). Acoustic and aerodynamic correlates of Korean stops and fricatives. *Journal of Phonetics 30*, 193–228.

Chung, Y.-H. (1991). *The lexical tone system of North Kyungsang Korean*. Ph. D. thesis, The Ohio State University.

Clayards, M. (2008). *The ideal listener: Making optimal use of acoustic-phonetic cues for word recognition*. Ph. D. thesis, University of Rochester.

Clayards, M., M. K. Tanenhaus, R. Aslin, and R. A. Jacobs (2008). Perception of speech reflects optimal use of probabilistic speech cues. *Cognition 108*, 804–809.

CMUdict (1998). CMU pronouncing dictionary version 0.6. http://www.speech.cs.cmu.edu/cgi-bin/cmudict.

Cohn, A. (1993). Nasalization in English: phonology or phonetics. *Phonology 10*, 43–81.

Cooper, F. S. (1953). Some instrumental aids to research on speech. In *Report on the Fourth Annual Round Table Meeting on Linguistics and Language Teaching*, pp. 46–53. Washington: Georgetown University Press.

Cooper, F. S., P. C. Delattre, A. M. Liberman, J. M. Borst, and L. G. Gerstman (1952). Some experiments on the perception of synthetic speech sounds. *Journal of the Acoustical Society of America 24*(6), 597–606.

Davis, M. H., I. S. Johnsrude, A. Hervais-Adelman, K. Taylor, and C. McGettigan (2005). Lexical information drives perceptual learning of distorted speech: evidence from the comprehension of noise-vocoded sentences. *Journal of Experimental Psychology: General 134*(2), 222–241.

Delattre, P. C., A. M. Liberman, F. S. Cooper, and L. J. Gerstman (1952). An experimental study of the acoustic determinants of vowel color: Observations on one- and two-formant vowels synthesized from spectrographic patterns. *Word 8*, 195–210.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B (Methodological) 39*(1), 1–38.

Devroye, L. (1986). *Non-Uniform Random Variate Generation*. New York: Springer Verlag. Online edition retrieved January 23, 2007 from http://cg.scs.carleton.ca/∼luc/rnbookindex.html.

Diehl, R. L. (2008). Acoustic and auditory phonetics: the adaptive design of speech sound systems. *Philosophical Transactions of the Royal Society 363*, 965–978.

Diehl, R. L. and R. Kluender (1989). On the objects of speech perception. *Ecological Psychology 1*, 121–144.

Dorman, M. F., M. Studdert-Kennedy, and L. J. Raphael (1977). Stop-consonant recognition: Release bursts and formant transitions as functionally equivalent, context-dependent cues. *Perception and Psychophysics 22*(2), 109–122.

Duanmu, S. (2000). *The Phonology of Standard Chinese*. Oxford: Oxford University Press.

Duda, R., P. Hart, and D. Stork (2000). *Pattern Classification*. New York: Wiley.

Dupoux, E. and K. Green (1997). Perceptual adjustment to highly compressed speech: effets of talker and rate changes. *Journal of Experimental Psychology: Human Perception and Performance 23*, 914–927.

Escudero, P. and P. Boersma (2004). Bridging the gap between L2 speech perception research and phonological theory. *Studies in Second Language Acquisition 26*, 551–585.

Estes, W. K. (1986). Array models for category learning. *Cognitive Psychology 18*(4), 500–549.

Feldman, N. H., T. L. Griffiths, and J. L. Morgan (2009). Learning phonetic categories by learning a lexicon. In *Proceedings of the 31st Annual Conference of the Cognitive Science Society*.

Ferguson, S. H. and D. Kewley-Port (2002). Vowel intelligibility in clear and conversational speech for normal-hearing and hearing-impaired listeners. *Journal of the Acoustical Society of America 112*(1), 259–271.

Flege, J. E. and R. Port (1981). Cross-language phonetic interference: Arabic to English. *Language and Speech 24*, 125–146.

Flemming, E. (2001). *Auditory Representations in Phonology*. New York: Garland Press.

Flemming, E. (2007). Commentary: Modeling listeners. In C. Fougeron, B. Kühnert, M. D'Imperio, and N. Vallée (Eds.), *Papers in Laboratory Phonology X: Variation, Phonetic Detail, and Phonological Modeling*. Berlin: Mouton de Gruyter.

Fraley, C. and A. E. Raftery (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association 97*(458), 611–631.

Fraley, C. and A. E. Raftery (2006). MCLUST Version 3 for R: Normal mixture modeling and model-based clustering. Technical Report 504, Department of Statistics, University of Washington.

Fraley, C. and A. E. Raftery (2007). Bayesian regularization for normal mixture estimation and model-based clustering. *Journal of Classification 24*, 155–181.

Francis, A. L., K. Baldwin, and H. C. Nusbaum (2000). Effects of training on attention to acoustic cues. *Perception & Psychophysics 62*, 1668–1680.

Francis, A. L., V. Ciocca, V. K. M. Wong, and J. K. L. Chan (2006). Is fundamental frequency a cue to aspiration in initial stops? *Journal of the Acoustical Society of America 120*(5), 2884–2895.

Francis, A. L., N. Kaganovich, and C. Driscoll-Huber (2008). Cue-specific effects of categorization training on the relative weighting of acoustic cues to consonant voicing in English. *Journal of the Acoustical Society of America 124*(2), 1234–1251.

Francis, A. L. and H. C. Nusbaum (2002). Selective attenion and the acquisition of new phonetic categories. *Journal of Experimental Psychology: Human Perception and Performance 28*, 349–366.

Gafos, A. (2002). A grammar of gestural coordination. *Natural Language and Linguistic Theory 20*, 269–337.

Gandour, J. T. (1974). Consonant types and tone in Siamese. *Journal of Phonetics 2*, 337–350.

Gandour, J. T. (1975). On the representation of tone in Siamese. In J. G. Harris and J. R. Chamberlain (Eds.), *Studies in Tai Linguistics in Honor of William J. Gedney*, pp. 170–195. Bangkok: Central Institute of English Language.

Garner, W. R. (1974). *The Processing of Information and Structure*. Hillsdale, NJ: Erlbaum.

Garofolo, J. S., L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, and V. Zue (1993). TIMIT acoustic-phonetic continuous speech corpus. http://www.ldc.upenn.edu/Catalog/LDC93S1.html.

Garrett, A. and K. Johnson (To appear). Phonetic bias in sound change. In A. C. L. Yu (Ed.), *Origins of Sound Patterns: Approaches to Phonologization*. Oxford: Oxford University Press.

Geisler, W. S. (1989). Sequential ideal-observer analysis of visual discrimination. *Psychological Review 96*, 267–314.

Giegerich, H. J. (1992). *English Phonology*. Cambridge: Cambridge University Press.

Goldinger, S. D. (1996). Words and voices: Episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition 22*(5), 1166–1183.

Goldsmith, J. (1976). *Autosegmental Phonology*. Ph. D. thesis, MIT. [Published by Garland Press, New York, 1979.].

Goldwater, S. (2006). *Nonparametric Bayesian models of lexical acquisition*. Ph. D. thesis, Brown University.

Goldwater, S., T. L. Griffiths, and M. Johnson (2009). A Bayesian framework for word segementation: Exploring the effects of context. *Cognition 112*(1), 21–54.

Gordon, P. C., J. L. Eberhardt, and J. G. Rueckl (1993). Attentional modulation of the phonetic significance of acoustic cues. *Cognitive Psychology 25*, 1–42.

Goudbeek, M. (2006). *The acquisition of auditory categories*. Ph. D. thesis, Radboud University, Nijmegen.

Goudbeek, M., A. Cutler, and R. Smits (2008). Supervised and unsupervised learning of multidimensionally varying non-native speech categories. *Speech Communication 50*, 109–125.

Green, D. and J. Swets (1966). *Signal Detection Theory and Psychophysics*. New York: Wiley.

Grünwald, P. (2007). *The Minimum Description Length Principle*. Cambridge, MA: MIT Press.

Guion, S. (1995). *Velar palatalization: coarticulation, perception, and sound change*. Ph. D. thesis, University of Texas at Austin.

Guion, S. G. (1998). The role of perception in the sound change of velar palatalization. *Phonetica 55*, 18–52.

Hagège, C. (2004). On categories, rules and interfaces in linguistics. *The Linguistic Review 21*, 257–276.

Hagège, C. and A.-G. Haudricourt (1978). *La phonologie panchronique*. Paris: Presses Universitaires de France.

Hajek, J. (1997). *Universals of Sound Change in Nasalization*. Oxford: Blackwell.

Halberstadt, A. K. (1998). *Heterogenous acoustic measurements and multiple classifiers for speech recognition*. Ph. D. thesis, Massachusetts Institute of Technology.

Halberstadt, A. K. and J. R. Glass (1997). Heterogeneous acoustic measurements for phonetic classification. In *EUROSPEECH 1997*, pp. 401–404.

Han, J.-I. (1996). *The phonetics and phonology of "tense" and "plain" consonants in Korean*. Ph. D. thesis, Cornell University, Ithaca, NY.

Han, M. S. and R. S. Weizman (1970). Acoustic features of Korean /P, T, K/, /p, t, k/ and /ph, th, kh/. *Phonetica 22*, 112–128.

Hansen, M. and B. Yu (1999). Bridging AIC and BIC: An MDL model selection criterion. In *Proceedings of the IEEE Information Theory Workshop on Detection, Estimation, Classification and Imaging*, Sante Fe, NM.

Hansen, M. and B. Yu (2001). Model selection and the Minimum Description Length principle. *Journal of the American Statistical Association 96*, 746–774.

Harrington, J., S. Palethorpe, and C. I. Watson (2000a). Does the Queen speak the Queen's English? *Nature 408*, 927–928.

Harrington, J., S. Palethorpe, and C. I. Watson (2000b). Monophthongal vowel changes in Received Pronunciation: An acoustic analysis of the Queen's Christmas broadcasts. *Journal of the International Phonetic Association 30*, 63–78.

Hasegawa-Johnson, M., J. Baker, S. Borys, K. Chen, E. Coogan, S. Greenberg, A. Juneja, K. Kirchhof, K. Livescu, S. Mohan, J. Muller, K. Sonmez, and T. Wang (2005). Landmark-based speech recognition: Report of the 2004 Johns Hopkins summer workshop. In *Proceedings of the IEEE Conference on Acoustic Speech Signal Processing 2005*, Volume 1, pp. 1213–1216.

Hastie, T., R. Tibshirani, and J. Friedman (2008). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.). New York: Springer Verlag.

Hayes, B., R. Kirchner, and D. Steriade (Eds.) (2004). *Phonetically-Based Phonology*. New York: Cambridge University Press.

Hewlett, N. (1988). Acoustic properties of /k/ and /t/ in normal and phonologically disorderd speech. *Clinical Linguistics and Phonetics 2*, 29–45.

Hickok, G. and D. Poeppel (2007). The cortical organization of speech processing. *Nature Reviews Neuroscience 8*, 393–402.

Hillenbrand, J. M., L. A. Getty, M. J. Clark, and K. Wheeler (1995). Acoustic characteristics of American English vowels. *Journal of the Acoustical Society of America 97*(5), 3099–3111.

Hirose, H., C. Y. Lee, and T. Ushijima (1974). Laryngeal control in Korean stop production. *Journal of Phonetics 2*, 145–152.

Hirose, H., H. S. Park, M. S. Hirohide Yoshioka, and H. Umeda (1981). An electromyographic study of laryngeal adjustments for the Korean stops. *Annual Bulletin of Research Institute of Logopedics and Phoniatrics 15*, 31–43.

Hockett, C. F. (1955). *A Manual of Phonology*, Volume 21-4 of *International Journal of American Linguistics*. Bloomington: Indiana University Publications.

Holt, L. L. and A. J. Lotto (2006). Cue weighting in auditory categorization: Implications for first and second language acquistion. *Journal of the Acoustical Society of America 119*, 3059–3071.

Hombert, J.-M. (1975). *Towards a theory of tonogenesis: an empirical, physiologically and perceptually-based account of the development of tonal contrasts in language*. Ph. D. thesis, University of California at Berkeley.

Hombert, J.-M. (1977a). Consonant types, vowel height, and tone in Yoruba. *Studies in African Linguistics 8*, 173–190.

Hombert, J.-M. (1977b). Development of tones from vowel height? *Journal of Phonetics 5*, 9–16.

Hombert, J.-M., J. J. Ohala, and W. G. Ewan (1979). Phonetic explanations for the development of tones. *Language 55*(1), 37–58.

Hong, S. (2007). *The phonetics of Daegu Korean lexical prosody*. Ph. D. thesis, The State University of New York at Buffalo.

House, A. S. and G. Fairbanks (1953). The influence of consonant environment on secondary acoustical characteristics of vowels. *Journal of the Acoustical Society of America 25*, 105–113.

Hume, E. and K. Johnson (2001). *The Role of Speech Perception in Phonology*. New York: Academic Press.

Hura, S. L., B. Lindblom, and R. L. Diehl (1992). On the role of perception in shaping phonological assimilation rules. *Language and Speech 35*(1–2), 59–72.

Hyman, L. M. (1972). Nasals and nasalization in Kwa. *Studies in African Linguistics 3*, 167–206.

Hyman, L. M. (1973). The role of consonant types in natural tonal assimilations. In L. M. Hyman (Ed.), *Consonant types and tone*, Volume 1 of *Southern California Occasional Papers in Linguistics*, pp. 151–179. Los Angeles: University of Southern California.

Hyman, L. M. (1976). Phonologization. In A. Juilland (Ed.), *Linguistic studies presented to Joseph H. Greenberg*, pp. 407–418. Saratoga: Anma Libri.

Hyman, L. M. (2000). Privative tone in Bantu. Paper presented at the Symposium on Tone, ILCAA, Tokyo, December 12–16, 2000.

Hyman, L. M. (2008). Enlarging the scope of phonologization. In *UC Berkeley Phonology Lab Annual Report (2008)*, pp. 382–408. Berkeley, CA: UC Berkeley Phonology Lab.

Hyman, L. M. (2009). How (not) to do phonological typology: the case of pitch-accent. *Language Sciences 2–3*, 213–238.

Iverson, P., V. Hazan, and K. Bannister (2005). Phonetic training with acoustic cue manipulations: A comparison of methods for teaching English /r/-/l/ to Japanese adults. *Journal of the Acoustical Society of America 118*(5), 3267–3278.

Jakobson, R. (1931). Prinzipien der historischen Phonologie. *Travaux du Cercle Linguistique de Prague 4*, 247–267.

Jakobson, R. (1941). *Kindersprache, Aphasie und allgemeine Lautgesetze*. Uppsala: Uppsala Universitets Arsskrift.

Jakobson, R., C. G. M. Fant, and M. Halle (1951). *Preliminaries to speech analysis*. Cambridge, MA: The MIT Press.

Johnson, K. (1997). Speech perception without speaker normalization. In J. W. Mullenix (Ed.), *Talker Variability in Speech Processing*, pp. 145–166. San Diego: Academic Press.

de Jong, K. (1991). An articulatory study of consonant-induced vowel duration changes in English. *Phonetica 48*, 1–17.

de Jong, K. (1995). The supraglottal articulation of prominence in English. *Journal of the Acoustical Society of America 97*, 491–504.

Jun, S.-A. (1993). *The phonetics and phonology of Korean prosody*. Ph. D. thesis, The Ohio State University.

Jun, S.-A. (1996). *The Phonetics and Phonology of Korean Prosody: Intonational Phonology and Prosodic Structure*. New York: Garland.

Juneja, A. (2004). *Speech recognition based on phonetic features and acoustic landmarks*. Ph. D. thesis, University of Maryland, College Park.

Kagaya, R. (1974). A fiberscopic and acoustic study of the Korean stops, affricates and fricatives. *Journal of Phonetics 2*, 161–180.

Kaji, S. (1996). Tone reversal in Tembo (Bantu J.57). *Journal of African Languages and Linguistics 17*, 1–26.

Kang, K.-H. and S. G. Guion (2008). Clear speech production of Korean stops: Changing phonetic targets and enhancement strategies. *Journal of the Acoustical Society of America 124*(6), 3909–3917.

Kaplan, A. (2010). *Phonology influenced by phonetics: the case of intervocalic lenition*. Ph. D. thesis, University of California at Santa Cruz.

Kavitskaya, D. (2002). *Compensatory Lengthening: Phonetics, Phonology, Diachrony*. New York: Routledge, Outstanding Dissertations in Linguistics.

Kessinger, R. H. and S. E. Blumstein (1997). Effects of speaking rate on voice-onset time in Thai, French, and English. *Journal of Phonetics 25*(2), 143–168.

Keyser, S. J. and K. N. Stevens (2001). Enhancement revisited. In M. Kenstowicz (Ed.), *Ken Hale: A life in language*, pp. 271–291. Cambridge, MA: MIT Press.

Keyser, S. J. and K. N. Stevens (2006). Enhancement and overlap in the speech chain. *Language 82*(1), 33–63.

Kim, C.-W. (1965). On the autonomy of the tensity feature in stop classification (with special reference to Korean stops). *Word 21*, 339–359.

Kim, M.-R. (2000). *Segmental and tonal interactions in English and Korean: a phonetic and phonological study*. Ph. D. thesis, University of Michigan.

Kim, M.-R., P. S. Beddor, and J. Horrocks (2002). The contribution of consonantal and vocalic information to the perception of Korean initial stops. *Journal of Phonetics 30*, 77–100.

Kim, M.-R. and S. Duanmu (2004). 'Tense' and 'lax' stops in Korean. *Journal of East Asian Linguistics 13*, 59–104.

Kim, M.-R. C. (1994). *Acoustic characteristics of Korean stops and perception of English stop consonants*. Ph. D. thesis, University of Wisconsin at Madison, Madison, WI.

Kim, Y.-K. (1975). *Korean Consonantal Phonology*. Seoul: Pagoda Press.

Kingston, J. (1985). *The phonetics and phonology of the timing of oral and glottal events*. Ph. D. thesis, University of California, Berkeley.

Kingston, J. (2005). The phonetics of Athabaskan tonogenesis. In S. Hargus and K. Rice (Eds.), *Athabaskan Prosody*. Amsterdam: John Benjamins.

Kingston, J. (2007). The phonetics-phonology interface. In P. de Lacy (Ed.), *The Handbook of Phonology*, pp. 401–434. Cambridge: Cambridge University Press.

Kingston, J. and R. L. Diehl (1994). Phonetic knowledge. *Language 70*, 419–454.

Kingston, J., R. L. Diehl, C. J. Kirk, and W. A. Castleman (2008). On the internal perceptual structure of distinctive features. *Journal of Phonetics 36*, 28–54.

Kiparsky, P. (1995). The phonological basis of sound change. In J. Goldsmith (Ed.), *The Handbook of Phonological Theory*, pp. 640–670. Oxford: Blackwell.

Kirby, S. (1999). *Function, Selection, and Innateness: The Emergence of Language Universals*. Oxford: Oxford University Press.

Kirby, S. and J. Hurford (2002). The emergence of linguistic structure: An overview of the Iterated Learning Model. In A. Cangelosi and D. Parisi (Eds.), *Simulating the Evolution of Language*, pp. 121–148. London: Springer Verlag.

Kirchner, R. (1998). *An effort-based account of consonant lenition*. Ph. D. thesis, University of California at Los Angeles.

Kirchner, R., R. K. Moore, and T.-Y. Chen (in press). Computing phonological generalization over real speech exemplars. *Journal of Phonetics*.

Kisseberth, C. and D. Odden (2003). Tone. In D. Nurse and G. Philippson (Eds.), *The Bantu Languages*, pp. 59–70. New York: Routledge.

Klatt, D. H. and L. C. Klatt (1990). Analysis, synthesis, and perception of voice quality variations among female and male talkers. *Journal of the Acoustical Society of America 87*(2), 820–857.

Ko, E.-S. (2003). The laryngeal effect in Korean: phonology or phonetics? In J. van de Weijer (Ed.), *Phonological Spectrum, Volume 1: Segmental structure*, pp. 171–191. Philadelphia: John Benjamins.

Kraljic, T. and A. G. Samuel (2006). Generalization in perceptual learning for speech. *Psychonomic Bulletin & Review 13*(2), 262–268.

Krauss, M. E. (2005 [1979]). Athabaskan tone. In S. Hargus and K. Rice (Eds.), *Athabaskan Prosody*. Amsterdam: John Benjamins.

Kruschke, J. K. (1992). Alcove: An exemplar-based connectionist model of category learning. *Psychological Review 99*(1), 22–44.

Kuhl, P. K. (1991). Human adults and human infants show a "perceptual magnet effect" for the prototypes of speech categories, monkeys do not. *Perception and Psychophysics 50*, 93–107.

Kuhl, P. K. (2004). Early language acquisition: Cracking the speech code. *Nature Reviews Neuroscience 5*, 831–843.

Kuhl, P. K., E. Stevens, A. Hayashi, T. Deguchi, S. Kiritani, and P. Iverson (2006). Infants show a facilitation effect for native language phonetic perception between 6 and 12 months. *Developmental Science 9*(2), F13–F21.

Kuhl, P. K., K. A. Williams, F. Lacerda, K. N. Stevens, and B. Lindblom (1992). Linguistic experience alters phonetic perception in infants by 6 months of age. *Science 255*, 606–608.

Kullback, S. and R. A. Leibler (1951). On information and sufficiency. *Annals of Mathematical Statitics 22*(1), 79–86.

Kwenzi-Mikala, J. (1980). Esquisse phonologique du punu. In F. Nsuka-Nkutsi (Ed.), *Eléments de description du punu*, pp. 7–18. Lyon: CRLS, Univeristé Lyon II.

Labov, W. (1990). The intersection of sex and social class in the course of linguistic change. *Language Variation and Change 2*, 205–254.

Labov, W. (1994). *Principles of Linguistic Change Vol. 1: Internal Factors*. Oxford: Oxford University Press.

Labov, W. (2001). *Principles of Linguistic Change Vol. 2: Social Factors*. Oxford: Oxford University Press.

Labov, W., M. Karen, and C. Miller (1991). Near-mergers and the suspension of phonemic contrast. *Language Variation and Change 3*, 33–74.

Labov, W., M. Yaeger, and R. Steiner (1972). *A quantitative study of sound change in progress*. U.S. Regional Survey: Philadelphia.

Lacerda, F. (1995). The perceptual-magnet effect: An emergent consequence of exemplar-based phonetic memory. In *Proceedings of the Thirteenth International Congress of the Phonetic Sciences*, Stockholm, pp. 140–147.

Ladd, R. (To appear). Phonetics in phonology. In J. Goldsmith, J. Riggle, and A. C. L. Yu (Eds.), *Handbook of Phonological Theory (2nd edition)*. Cambridge: Wiley-Blackwell.

Ladefoged, P. (1967). *Linguistic Phonetics*. Number 6 in UCLA Working Papers in Phonetics. Los Angeles: UCLA.

Ladefoged, P. (1973). The features of the larynx. *Journal of Phonetics 1*, 73–83.

Lahiri, A., H. Schriefers, and C. Kuijpers (1987). Contextual neutralization of vowel length: Evidence from Dutch. *Phonetica 44*, 91–102.

Lass, R. (1980). *On Explaining Language Change*. Cambridge: Cambridge University Press.

Laver, J. (1980). *The Phonetic Description of Voice Quality*. Cambridge University Press: Cambridge.

Lee, C. Y. and T. S. Smith (1972). Oral and direct subglottal pressure in Korean stops. *Journal of the Acoustical Society of America 51*(1), 102.

Lee, J., T. K. Perrachione, T. M. Dees, and P. C. M. Wong (2007). Differential effects of stimulus variability and learners' pre-existing pitch perception ability in lexical tone learning by native English speakers. In *Proceedings of the 16th International Congress of Phonetic Sciences*, Saabrücken, pp. 717–720.

Leer, J. A. (1999). Tonogenesis in Athapaskan. In S. Kaji (Ed.), *Cross-linguistic studies of tonal phenomena: Tonogenesis, typology, and related topics*, pp. 37–66. Tokyo: Institute of the Study of the Languages and Cultures of Asia and Africa, Tokyo University of Foreign Studies.

Lehiste, I. and G. E. Peterson (1961). Some basic considerations in the analysis of intonation. *Journal of the Acoustical Society of America 33*, 419–425.

Li, F., A. Menon, and J. B. Allen (2010). A psychoacoustic method to find the perceptual cues of stop consonants in natural speech. *Journal of the Acoustical Society of America 127*(4), 2599–2610.

Liberman, A. M. and F. S. Cooper (1972). In search of the acoustic cues. In A. Valdman (Ed.), *Papers on Linguistics and Phonetics to the Memory of Pierre Delattre*, pp. 329–338. The Hague: Mouton.

Liberman, A. M., F. S. Cooper, D. P. Shankweiler, and M. Studdert-Kennedy (1967). Perception of the speech code. *Psychological Review 74*(6), 431–461.

Liberman, A. M., P. C. Delattre, and F. S. Cooper (1952). The role of selected stimulus variables in the perception of unvoiced stop consonants. *American Journal of Psychology 65*, 497–516.

Liberman, A. M., K. S. Harris, H. S. Hoffmann, and B. C. Griffith (1957). The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology 54*(5), 358–368.

Liberman, M. (2000). The "lexical contract": Modeling the emergence of word pronunciations. Talk given in the IRCS Colloquium Series, Sept. 22, 2000. http://www.ldc.upenn.edu/myl/abm/index.html, retrieved on Sept. 22, 2010.

Liljencrants, J. and B. Lindblom (1972). Numerical simulation of vowel quality systems: the role of perceptual contrast. *Language 48*, 839–862.

Lin, Y. (2005). *Learning features and segments from waveforms: a statistical model of early phonological acquisition*. Ph. D. thesis, UCLA.

Lindblom, B. (1990). Explaining phonetic variation: a sketch of the H & H theory. In *Speech Production and Speech Modeling*, pp. 403–439. Dordrecht: Kluwer.

Lindblom, B., S. Guion, S. Hura, S.-J. Moon, and R. Willerman (1995). Is sound change adaptive? *Rivista di Linguistica 7*(1), 5–37.

Lippmann, R. P. (1996). Accurate consonant perception without mid-frequency speech energy. *IEEE Transations on Speech and Audio Processing 4*(1), 66–69.

Lisker, L. (1978). Rapid vs. rabid: a catalogue of acoustic features that may cue the distinction. Technical report.

Lisker, L. and A. Abramson (1964). A cross-language study of voicing in initial stops: acoustical measurements. *Word 20*, 384–422.

Lisker, L. and A. Abramson (1970). The voicing dimension: some experiments in comparative phonetics. In *Proceedings of the Sixth International Congress of Phonetic Sciences*, Prague, pp. 563–567. Academia Publishing House of the Czechoslovak Academy of Sciences.

Liu, S. A. (1995). *Landmark detection for distinctive feature-based speech recognition*. Ph. D. thesis, Massachusetts Institute of Technology.

Luce, R. D. (1959). *Individual Choice Behavior*. New York: Wiley.

Luce, R. D. (1963). Detection and recognition. In R. D. Luce, R. R. Bush, and E. Galanter (Eds.), *Handbook of Mathematical Psychology*, Volume I, pp. 103–189. New York: Wiley.

Macken, M. A. and D. Barton (1980). The acquisition of the voicing contrast in English: a study of voice onset time in word-initial stop consonants. *Journal of Child Language 7*, 41–74.

Macmillan, N. A. and C. D. Creelman (1981). *Detection Theory: A User's Guide*. New York: Cambridge University Press.

Maddieson, I. (1976). Tone reversal in Chiluba: a new theory. In L. M. Hyman (Ed.), *Studies in Bantu tonology (SCOPIL #3)*, pp. 141–165. Los Angeles: University of Southern California.

Maddox, W. T. and F. G. Ashby (1998). Selective attention and the formation of linear decision bounds. commentary on mckinley and nosofsky (1996). *Journal of Experimental Psychology: Learning, Memory and Cognition 24*, 301–321.

Magnuson, J. S. and H. C. Nusbaum (2007). Acoustic differences, listener expectations, and the perceptual accommodation of talker variability. *Journal of Experimental Psychology: Human Perception and Performance 33*(2), 391–409.

Mann, V. A. and B. H. Repp (1980). Influence of vocalic context on the [s]-[ʃ] distinction. *Perception and Psychophysics 28*, 213–228.

Martin, A. W. (2007). *The evolving lexicon*. Ph. D. thesis, UCLA.

Martinet, A. (1952). Function, structure, and sound change. *Word 8*(1), 1–32.

Massaro, D. W. (1987). *Speech Perception by Eye and Ear: A Paradigm for Psychological Inquiry*. Hillsdale, NJ: Erlbaum.

Massaro, D. W. and G. C. Oden (1980). Evaluation and integration of acoustic features in speech perception. *Journal of the Acoustical Society of America 67*, 996–1013.

Matisoff, J. A. (1973). Tonogenesis in Southeast Asia. In L. Hyman (Ed.), *Consonant Types and Tone*, pp. 71–95. USC: SCOPIL.

Maye, J., J. F. Werker, and L. Gerken (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition 82*(3), B101–B111.

McKinley, S. C. and R. M. Nosofsky (1995). Investigations of exemplar and decision bound models in large, ill-defined category structures. *Journal of Experimental Psychology: Human Perception and Performance 21*(1), 128–148.

McLachlan, G. J. and D. Peel (2000). *Finite Mixture Models*. New York: Wiley.

McMurray, B., R. N. Aslin, and J. C. Toscano (2009). Statistical learning of phonetic categories: insights from a computational approach. *Developmental Science 12*(3), 369–378.

McMurray, B., M. K. Tanenhaus, and R. N. Aslin (2002). Gradient effects of within-category phonetic varition on lexical access. *Cognition 86*, B33–B42.

Melnykov, V. and R. Maitra (2010). Finite mixture models and model-based clustering. *Statistics Surveys 4*, 80–116.

Miller, J. L. and L. E. Volaitis (1989). Effect of speaking rate on the perceptual structure of a phonetic category. *Perception & Psychophysics 46*(6), 505–512.

Moon, S.-J. and B. Lindblom (1994). Interaction between duration, contextm and speaking style in English stressed vowels. *Journal of the Acoustical Society of America 109*(3), 1181–1196.

Moreton, E. (2002). Structural constraints in the perception of English stop-sonorant clusters. *Cognition 84*, 55–71.

Moreton, E. (2008). Analytic bias and phonological typology. *Phonology 25*, 83–127.

Morrison, G. S. (2005). An appropriate metric for cue weighting in L2 speech perception: Response to Escudero and Boersma (2004). *Studies in Second Language Acquisition 27*(4), 597–606.

Nash, J. A. (1994). Underlying low tones in Ruwund. *Studies in African Linguistics 23*, 223–278.

Nearey, T. (1997). Speech perception as pattern recognition. *Journal of the Acoustical Society of America 101*(6), 3241–3254.

Nearey, T. and J. T. Hogan (1986). Phonological contrast in experimental phonetics: relating distributions of production data to perceptual categorization curves. In J. J. Ohala and J. J. Jaeger (Eds.), *Experimental Phonology*, pp. 141–162. Orlando: Academic Press.

Newman, R. S., S. A. Clouse, and J. L. Burnham (2001). The perceptual consequences of within-talker variability in fricative production. *Journal of the Acoustical Society of America 109*(3), 1181–1196.

Nittrouer, S. (2004). The role of temporal and dynamic signal components in the perception of syllable-final stop voicing by children and adults. *Journal of the Acoustical Society of America 115*(4), 1777–1790.

Niyogi, P. (2006). *The Computational Nature of Language Learning and Evolution*. Cambridge: The MIT Press.

Niyogi, P. and R. C. Berwick (1995). The logical problem of language change. Technical Report AI Memo 1516 / CBCL Paper 115, Massachusetts Institute of Technology. MIT AI Laboratory and Center for Biological and Computational Learning, Department of Brain and Cognitive Sciences.

Niyogi, P. and R. C. Berwick (1996). A language learning model for finite parameter spaces. *Cognition 61*(1-2), 161–193.

Niyogi, P. and R. C. Berwick (2009). The proper treatment of language acquisition and change in a population setting. *Proceedings of the National Academy of Sciences 106*(25), 10124–10129.

Norris, D. (2006). The Bayesian reader: explaining word recgonition as an optimal Bayesian decision process. *Psychological Review 113*, 327–257.

Norris, D. and J. M. McQueen (2008). Shortlist B: a Bayesian model of continuous speech recognition. *Psychological Review 115*(2), 357–395.

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General 115*(1), 39–57.

Nosofsky, R. M. (1990). Relations between exemplar-similarity and likelihood models of classification. *Journal of Mathematical Psychology 34*, 393–418.

Nosofsky, R. M. (1998). Selective attention and the formation of linear decision boundaries: Reply to maddox and ashby (1998). *Journal of Experimental Psychology: Human Perception and Performance 24*(1), 322–339.

Nosofsky, R. M. and S. R. Zaki (2002). Exemplar and prototype models revisited: Response strategies, selective attention, and stimulus generalization. *Journal of Experimental Psychology: Learning, Memory, and Cognition 28*(5), 924–940.

Nusbaum, H. C. and J. Magnuson (1997). Talker normalization: Phonetic constancy as a cognitive process. In K. Johnson and J. W. Mullenix (Eds.), *Talker Variability in Speech Processing*, pp. 109–132. San Diego: Academic Press.

Nusbaum, H. C. and E. C. Schwab (1986). The role of attention and active processing in speech perception. In E. C. Schwab and H. C. Nusbaum (Eds.), *Pattern Recognition by Humans and Machines*, pp. 113–157. San Diego: Academic Press.

Nygaard, L. C. and D. B. Pisoni (1995). Speech perception: New directions in research and theory. In J. L. Miller and P. D. Eimas (Eds.), *Handbook of Perception and Cognition, Volume II: Speech, Language and Communication*, pp. 63–96. New York: Academic Press.

Oden, G. C. and D. W. Massaro (1978). Integration of featural information in speech perception. *Psychological Review 85*, 172–191.

Ohala, J. J. (1973). The physiology of tone. In L. Hyman (Ed.), *Consonant Types and Tone*, pp. 1–14. USC: SCOPIL.

Ohala, J. J. (1981a). The listener as a source of sound change. In C. Masek, R. Hendrick, and M. Miller (Eds.), *CLS 17-2: Papers from the Parasession on Language and Behavior*, pp. 178–203. Chicago: Chicago Linguistic Society.

Ohala, J. J. (1981b). Speech timing as a tool in phonology. *Phonetica 38*, 204–217.

Ohala, J. J. (1983). The origin of sound patterns in vocal tract constraints. In P. MacNeilage (Ed.), *The Production of Speech*, pp. 189–216. New York: Springer Verlag.

Ohala, J. J. (1989). Sound change is drawn from a pool of synchronic variation. In L. Breivik and E. Jahr (Eds.), *Language change: contributions to the study of its causes*, Trends in Linguistics, Studies and Monographs No. 43, pp. 173–198. Berlin: Mouton de Gruyter.

Ohala, J. J. (1990). There is no interface between phonology and phonetics: a personal view. *Journal of Phonetics 18*, 153–171.

Ohala, J. J. (1993a). The phonetics of sound change. In C. Jones (Ed.), *Historical Linguistics: Problems and Perspectives*, pp. 237–278. London: Longman.

Ohala, J. J. (1993b). Sound change as nature's speech perception experiment. *Speech Communication 13*, 155–161.

Ohde, R. N. and K. N. Stevens (1983). Effect of burst amplitude on the perception of stop consonant place of articulation. *Journal of the Acoustical Society of America 74*, 706–714.

Paul, H. (1880 [1889]). *Prinzipien der Sprachgeschichte [Principles of the history of language]*. New York: Macmillan & Co. Translated by H. A. Strong.

Pearson, K. (1894). Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society A 185*, 71–110.

Picheny, M. A., N. I. Durlach, and L. D. Braida (1986). Speaking clearly for the hard of hearing ii: Acoustic characteristics of clear and conversational speech. *Journal of Speech and Hearing Research 29*, 434–446.

Pierrehumbert, J. (1980). *The phonology and phonetics of English intonation*. Ph. D. thesis, Massachusetts Institute of Technology.

Pierrehumbert, J. (2001). Exemplar dynamics: word frequency, lenition, and contrast. In J. Bybee and P. Hopper (Eds.), *Frequency effects and the emergence of linguistic structure*, pp. 137–157. Amsterdam: John Benjamins.

Pierrehumbert, J. (2002). Word-specific phonetics. In C. Gussenhoven and N. Warner (Eds.), *Papers in Laboratory Phonology VII*, pp. 101–139. Berlin: Mouton de Gruyter.

Pisoni, D. B. (1992). Some comments on invariance, variability and perceptual normalization in speech perception. In *Proceedings 1992 International Conference on Spoken Language Processing, Banff, Canada, Oct. 12-16, 1992*, pp. 587–590.

Pisoni, D. B. and J. Tash (1974). Reaction times to comparisons within and across phonetic categories. *Perception and Psychophysics 15*(2), 285–290.

Plauché, M. (2001). *Acoustic cues in the directionality of stop consonant confusions*. Ph. D. thesis, University of California, Berkeley.

Plauché, M., C. Delogu, and J. J. Ohala (1997). Asymmetries in consonant confusion. In *EUROSPEECH-1997*, pp. 2187–2190.

Port, R. and P. Crawford (1989). Pragmatic effects on neutralization rules. *Journal of Phonetics 16*, 257–282.

Port, R. F. and J. Dalby (1982). Consonant/vowel ratio as a cue for voicing in English. *Perception & Psychophysics 32*(2), 141–152.

Posner, M. I. and S. W. Keele (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology 77*, 353–363.

Pothos, E. M. and T. M. Bailey (2009). Predicting category intuitiveness with the rational model, the simplicity model, and the Generalized Context Model. *Journal of Experimental Psychology: Learning, Memory, and Cognition 35*, 1062–1080.

Pothos, E. M. and N. Chater (2002). A simplicity principle in unsupervised human categorization. *Cognitive Science 26*, 393–343.

Pothos, E. M. and J. Close (2008). One or two dimensions in spontaneous classification: A simplicity approach. *Cognition 107*, 581–602.

Pouplier, M. (2010). Re-examining the contribution of articulatory effort to sound change. Paper presented at the Workshop on Sound Change, Barcelona, 21 October 2010.

Przezdziecki, M. (2005). *Vowel harmony and coarticulation in three dialects of Yoruba: Phonetics determining phonology*. Ph. D. thesis, Cornell University.

Pulleyblank, E. G. (1991). *Lexicon of reconstructed pronunciation in Early Middle Chinese, Late Middle Chinese, and Early Mandarin*. Vancouver: UBC Press.

Rasmussen, C. E. (2000). The infinite Gaussian mixture model. In S. A. Solla, T. K. Leen, and K. R. M uller (Eds.), *Advances in Neural Information Processing Systems*, Volume 12, pp. 554–560. Cambridge, MA: MIT Press.

Recasens, D., M. D. Pallerès, and J. Fontdevila (1997). A model of lingual coarticulation based on articulatory constraints. *Journal of the Acoustical Society of America 102*(1), 544–561.

Reed, S. K. (1972). Pattern recognition and categorization. *Cognitive Psychology 3*, 383–407.

Remez, R. E., P. E. Rubin, D. B. Pisoni, and T. D. Carrell (1981). Speech perception without traditional speech cues. *Science 212*, 947–950.

Repp, B. H. (1982). Phonetic trading relations and context effects: new experimental evidence for a speech mode of perception. *Psychological Bulletin 92*, 81–110.

Rissanen, J. (1978). Modeling by shortest data description. *Automatica 14*, 445–471.

Rogers, H. (2005). *Writing Systems: A Linguistic Approach*. Oxford: Blackwell.

Rosch, E. H. (1973). Natural categories. *Cognitive Psychology 4*, 328–350.

Rosseel, Y. (2002). Mixture models of categorization. *Journal of Mathematical Psychology 46*(2), 178–210.

Sampson, R. (1999). *Nasal Vowel Evolution in Romance*. Oxford: Blackwell.

Samuel, A. G. (1982). Phonetic prototypes. *Perception and Psychophysics 31*, 307–314.

Schwarz, G. E. (1978). Estimating the dimension of a model. *Annals of Statistics 6*(2), 461–464.

Scobbie, J. M., F. Gibbon, W. J. Hardcastle, and P. Fletcher (2000). Covert contrast as a stage in the acquisition of phonetics and phonology. In M. Broe and J. Pierrehumbert (Eds.), *Papers in Laboratory Phonology V: Language Acquisition and the Lexicon*, pp. 194–207. Cambridge: Cambridge University Press.

Shannon, C. E. and W. Weaver (1949). *A Mathematical Theory of Communication*. Urbana: University of Illinois Press.

Shannon, R. V., F.-G. Zeng, V. Kamath, J. Wygonski, and M. Ekelid (1995). Speech recognition with primarily temporal cues. *Science 270*(5234), 303–304.

Shepard, R. N. (1957). Stimulus and response generalization: A stochastic model relating generalization to distance in psychological space. *Psychometrika 22*, 325–345.

Shepard, R. N. (1958). Stimulus and response generalization: Tests of a model relating generalization to distance in psychological space. *Journal of Experimental Psychology 55*, 509–523.

Silva, D. J. (1992). *The phonetics and phonology of stop lenition in Korean.* Ph. D. thesis, Cornell University.

Silva, D. J. (1993). A phonetically based analysis of [voice] and [fortis] in Korean. In P. M. Clancy (Ed.), *Japanese/Korean Linguistics, Vol. 2*, pp. 164–174. Stanford: CSLI.

Silva, D. J. (2006a). Acoustic evidence for the emergence of tonal contrast in contemporary Korean. *Phonology 23*, 287–308.

Silva, D. J. (2006b). Variation in voice onset time for Korean stops: a case for recent sound change. *Korean Linguistics 13*, 1–16.

Silverman, D. (2006). *A Critical Introduction to Phonology: Of Sound, Mind, and Body.* London: Continuum.

Silverman, D. (2010). Neutralization and anti-homophony in Korean. *Journal of Linguistics 46*, 453–482.

Singh, R., B. Raj, and R. M. Stern (2000). Structured redefinition of sound units for improved speech recognition. In *Proc. 6th Intl. Conf. on Speech and Language Processing (ICSLP 2000)*, Beijing, China.

Smiljanić, R. and A. R. Bradlow (2008). Stability of temporal contrasts across speaking styles in Croatian and English. *Journal of Phonetics 36*, 91–113.

Smith, J. D. and J. P. Minda (2000). Thirty categorization results in search of a model. *Journal of Experimental Psychology: Learning, Memory, and Cognition 25*, 3–27.

Smits, R. (1996). A pattern-recognition-based framework for research on phonetic perception. In *Speech Hearing and Language Work in Progress*, Volume 9, pp. 195–229. London: University College Department of Phonetics and Linguistics.

Smits, R., J. Sereno, and A. Jongman (2006). Categorization of sounds. *Journal of Experimental Psychology: Human Perception and Performance 32*(3), 733–754.

Smits, R., L. ten Bosch, and R. Collier (1996a). Evaluation of various sets of acoustic cues for the perception of prevocalic stop consonants. I. Perception experiment. *Journal of the Acoustical Society of America 100*, 3852–3864.

Smits, R., L. ten Bosch, and R. Collier (1996b). Evaluation of various sets of acoustic cues for the perception of prevocalic stop consonants. II. Modeling and evaluation. *Journal of the Acoustical Society of America 100*, 3852–3864.

Solé, M.-J. (1992). Phonetic and phonological processes: the case of nasalization. *Language and Speech 35*(1–2), 29–43.

Solé, M.-J. (1995). Spatio-temporal patterns of velo-pharyngeal action in phonetic and phonological nasalization. *Language and Speech 38*(1), 1–23.

Solé, M.-J. (2003). Is variation encoded in phonology? In *Proceedings of the Fifteenth International Congress of the Phonetic Sciences*, Barcelona, pp. 289–292.

Solé, M.-J. (2007). Controlled and mechanical properties in speech. In M.-J. Solé, P. S. Beddor, and M. Ohala (Eds.), *Experimental Approaches to Phonology*, pp. 302–321. Oxford: Oxford University Press.

Sonderegger, M. and P. Niyogi (2010). Combining data and mathematical models of language change. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, Uppsala, Sweden, pp. 1019–1029.

van Spaandonck, M. (1971). On the so-called reversive tonal system of Chiluba. *Studies in African Linguistics 2*, 131–144.

Steriade, D. (1997). Phonetics in phonology: the case of laryngeal neutralization. *UCLA Working Papers in Linguistics 3*, 25–146.

Stevens, K. (2000). *Acoustic Phonetics (Current Studies in Linguistics)*. Cambridge, MA: The MIT Press.

Stevens, K. N. (1985). Evidence for the role of acoustic boundaries in the perception of speech sounds. In V. E. Fromkin (Ed.), *Phonetic Linguistics - Essays in Honor of Peter Ladefoged*, pp. 243–255. Orlando, FL: Academic Press.

Stevens, K. N. (1995). Applying phonetic knowledge to lexical access. In *Proceedings of Eurospeech-95*, Volume 1, pp. 3–11.

Stevens, K. N. and S. E. Blumstein (1981). The search for invariant acoustic correlates of phonetic features. In P. D. Eimas and J. L. Miller (Eds.), *Perspectives on the Study of Speech*, pp. 1–39. Hillsdale, NJ: Lawrence Erlbaum.

Summerfield, Q. (1975). Aerodynamics versus mechanics in the control of voicing onset in consonant-vowel syllables. In *Speech Perception (No. 4)*. Belfast: Queen's University, Department of Psychology.

Surendran, D. and P. Niyogi (2003). Measuring the usefulness (functional load) of phonological contrasts. Technical Report TR-2003-12, Department of Computer Science, University of Chicago, Chicago.

Surendran, D. and P. Niyogi (2006). Quantifying the functional load of phonemic oppositions, distinctive features, and suprasegmentals. In O. N. Thomsen (Ed.), *Competing Models of Linguistic Change: Evolution and beyond*, pp. 43–58. Amsterdam: John Benjamins.

Suwilai, P. (2001). Tonogenesis in Khmu dialects of SEA. *Mon-Khmer Studies 31*, 47–56.

Svantesson, J.-O. (1983). *Kammu Phonology and Morphology*. Lund: CWK Gleerup.

Teh, Y. W., M. I. Jordan, M. J. Beal, and D. M. Blei (2006). Hierarchical Dirichlet processes. *Journal of the American Statistical Association 101*, 1566–1581.

Toscano, J. C. and B. McMurray (2008). Using the distributional statistics of speech sounds for weighting and integrating acoustic cues. In *Proceedings of the Cognitive Science Society*, Mahwah, NJ. Erlbaum.

Toscano, J. C. and B. McMurray (2010). Cue integration with categories: weighting acoustic cues in speech using unsupervised learning and distributional statistics. *Cognitive Science 34*, 434–464.

Vallabha, G. K., J. L. McClelland, F. Pons, J. F. Werker, and S. Amano (2007). Unsupervised learning of vowel categories from infant-directed speech. *Proceedings of the National Academy of Sciences 104*(33), 13273–13278.

Wang, H. (2006). Nearest neighbors by neighborhood counting. *IEEE Transactions on Pattern Analysis and Machine Intelligence 28*(6), 942–953.

Wang, W. S.-Y. and C. J. Fillmore (1961). Intrinsic cue and consonant perception. *Journal of Speech & Hearing Research 4*, 130–136.

Wang, W. S.-Y., J. Ke, and J. W. Minett (2004). Computational studies of language evolution. In C.-R. Huang and W. Lenders (Eds.), *Computational Linguistics and Beyond: Frontiers in Linguistics 1, Languages and Linguistics Monograph Series B*, pp. 65–108. Institute of Linguistics: Academia Sinica.

Warner, N., A. Jongman, J. Sereno, and R. Kemps (2004). Incomplete neutralization and other sub-phonemic durational differences in production and perception: evidence from Dutch. *Journal of Phonetics 32*, 251–276.

Wedel, A. B. (2004). Category competition drives contrast maintenance within an exemplar-based production/perception loop. In J. Goldsmith and R. Wicentowski (Eds.), *Proceedings of the Seventh Meeting Meeting of the ACL Special Interest Group on Computational Phonology*, pp. 1–10.

Wedel, A. B. (2006). Exemplar models, evolution and language change. *The Linguistic Review 23*, 247–274.

Werker, J. F., J. H. V. Gilbert, K. Humphrey, and R. C. Tees (1981). Developmental aspects of cross-language speech perception. *Child Development 52*, 349–355.

Werker, J. F. and R. C. Tees (1984). Phonemic and phonetic factors in adult cross-language speech perception. *Journal of the Acoustical Society of America 75*, 1866–1878.

Wong, P. C. M. and T. K. Perrachione (2007). Learning pitch patterns in lexical identification by native English-speaking adults. *Applied Psycholinguistics 28*(4), 565–585.

Wright, J. (2007). *Laryngeal contrasts in Seoul Korean*. Ph. D. thesis, University of Pennsylvania, Philadelphia, PA.

Wright, R., S. Hargus, and K. Davis (2002). On the categorization of ejectives: Data from Witsuwit'en. *Journal of the International Phonetic Association 32*, 43–77.

Yu, A. C. L. (2007). Understanding near mergers: the case of morphological tone in Cantonese. *Phonology 24*, 187–214.

Yu, A. C. L. (To appear). Mergers and neutralization. In M. van Oostendorp, C. Ewen, E. Hume, and K. Rice (Eds.), *Blackwell Companion to Phonology*. Cambridge: Wiley-Blackwell.

Zipf, G. K. (1949). *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology*. New York: Hafner.