Indicators of Allophony and Phonemehood

Luc Boruta

UNIVERSITÉ PARIS–DIDEROT (PARIS 7)
ÉCOLE DOCTORALE INTERDISCIPLINAIRE EUROPÉENNE FRONTIÈRES DU VIVANT

DOCTORAT NOUVEAU RÉGIME — SCIENCES DU VIVANT

LUC BORUTA

INDICATORS *of* ALLOPHONY & PHONEMEHOOD

INDICATEURS D'ALLOPHONIE ET DE PHONÉMICITÉ

Thèse sous la direction de
Benoît CRABBÉ & Emmanuel DUPOUX

Soutenue le 26 septembre 2012

JURY

Mme Martine ADDA-DECKER, rapporteuse
Mme Sharon PEPERKAMP, rapporteuse
M. John NERBONNE, examinateur
M. Benoît CRABBÉ, directeur de thèse
M. Emmanuel DUPOUX, directeur de thèse

CONTENTS

ACKNOWLEDGEMENTS

I had wonderful support and encouragement while preparing this dissertation. First and foremost, Benoît Crabbé—my advisor and longtime *maître à penser*—provided invaluable feedback, expert insight, and inspirational discussion during the six years I spent working my way up through the computational linguistics program at Paris Diderot. His wise selection of textbooks, Belgian beers, and programming tips never proved wrong.

Emmanuel Dupoux—my other advisor—gave essential critical feedback throughout my graduate studies. He also provided me with a large quantity of excellent data on which to work.

Sharon Peperkamp, Martine Adda-Decker, and John Nerbonne graciously agreed—and on short notice—to carefully proofread my work with scientific rigour and a critical eye. I thank them for their enthusiastic and constructive comments.

François Taddei and Samuel Bottani welcomed me into a top-notch graduate school and an even more exciting and accepting community of knowledge addicts. I learned a great deal in their company, including new ways to learn. I owe them both a tremendous amount of gratitude.

Charlotte Roze—my partner in misery—has been a constant source of support in and out of the academic world. She knows that we can have it all, and everything will be alright.

Finally, a big thank goes out to Benoît Healy who has supported me beyond measure, and who spent hours considerately chasing the Frenglish in this dissertation. I know his brain hurts.

So, come up to the lab, and see what's on the slab...

ABSTRACT

Although we are only able to distinguish between a finite, small number of sound categories—i.e. a given language's phonemes—no two sounds are actually identical in the messages we receive. Given the pervasiveness of sound-altering processes across languages—and the fact that every language relies on its own set of phonemes—the question of the acquisition of allophonic rules by infants has received a considerable amount of attention in recent decades. How, for example, do English-learning infants discover that the word forms [kæt] and [kʰæt] refer to the same animal species (i.e. *cat*), whereas [kæt] and [bæt] (i.e. *cat* ∼ *bat*) do not? What kind of cues may they rely on to learn that [sɪŋkɪŋ] and [θɪŋkɪŋ] (i.e. *sinking* ∼ *thinking*) can not refer to the same action? The work presented in this dissertation builds upon the line of computational studies initiated by Peperkamp et al. (2006), wherein research efforts have been concentrated on the definition of sound-to-sound dissimilarity measures indicating which sounds are realizations of the same phoneme. We show that solving Peperkamp et al.'s task does not yield a full answer to the problem of the discovery of phonemes, as formal and empirical limitations arise from its pairwise formulation. We proceed to circumvent these limitations, reducing the task of the acquisition of phonemes to a partitioning-clustering problem and using multidimensional scaling to allow for the use of individual phones as the elementary objects. The results of various classification and clustering experiments consistently indicate that effective indicators of allophony are not necessarily effective indicators of phonemehood. Altogether, the computational results we discuss suggest that allophony and phonemehood can only be discovered from acoustic, temporal, distributional, or lexical indicators when—on average—phonemes do not have many allophones in a quantized representation of the input.

❝In the beginning it was too far away for Shadow to focus on. Then it became a distant beam of hope, and he learned how to tell himself 'this too shall pass'.❞

— Neil Gaiman, in *American Gods*

CHAPTER 1
INTRODUCTION

1.1 Problem at hand

Sounds are the backbone of daily linguistic communication. Not all natural languages have written forms; even if they do, speech remains the essential medium on which we rely to deliver information—mostly because it is always at one's disposal, particularly when other communication media are not. Furthermore, verbal communication often appears to be effortless, even to children with relatively little experience using their native language. We are hardly conscious of the various linguistic processes at stake in the incessant two-way coding that transforms our ideas into sounds, and vice versa.

The elementary sound units that we use as the building blocks of our vocalized messages are, however, subject to a considerable amount of variability. Although we are only able to distinguish between a finite, small number of sound categories—that linguists refer to as a given language's phonemes—no two sounds are actually identical in the messages we receive. Sounds do not only vary because everyone's voice is unique and, to some extent, characterized by one's gender, age, or mood, they also vary because each language's grammar comprises a substantial number of so called allophonic rules that constrain the acoustic realization of given phonemes in given contexts. In English, for example, the first /t/ in *tomato* is—beyond one's control—different from the last: whereas the first, most likely transcribed as [t^h], is followed by a burst of air, the last is a plain [t] sound. This discrepancy emblematizes a feature of the grammar of English whereby the consonants /p/, /t/, and /k/ are followed by a burst of air—an aspiration—when they occur as the initial phoneme of a word or of a stressed syllable. Given the pervasiveness of such sound-altering processes across languages—and the fact that every language relies on its own set of phonemes—the question of the acquisition of allophonic rules by infants has received a considerable amount of attention in recent decades. How, for example, do English-learning infants discover that the word forms [kæt] and [kat] refer to the same animal species (i.e. *cat*), whereas [kæt] and [bæt] (i.e. *cat* $\sim$ *bat*) do not ?

1.2 Motivation and contribution

Broadly speaking, research on early language acquisition falls into one of two categories: behavioral experiments and computational experiments. The purpose of this dissertation is to present work carried out for the computational modeling of the acquisition of allophonic rules by infants, with a focus on the examination of the relative informativeness of different types of cues on which infants may rely—viz. acoustic, temporal, distributional, and lexical cues.

The work presented in this dissertation builds upon the line of computational studies initiated by Peperkamp et al. (2006), wherein research efforts have been concentrated on the definition of sound-to-sound dissimilarity measures indicating which sounds are realizations of the same

phoneme. The common hypothesis underlying this body of work is that infants are able to keep track of—and rely on—such dissimilarity judgments in order to eventually cluster similar sounds into phonemic categories. Because no common framework had yet been proposed to systematically define and evaluate such empirical measures, our focus throughout this study was to introduce a flexible formal apparatus allowing for the specification, combination, and evaluation of what we refer to as indicators of allophony. In order to discover the phonemic inventory of a given language, the learning task introduced by Peperkamp et al. consists in predicting—for every possible pair of sounds in a corpus, given various indicators of allophony—whether or not two sounds are realizations of the same phoneme. In this dissertation, we show that solving this task does not yield a full answer to the problem of the discovery of phonemes, as formal and empirical limitations arise from its pairwise formulation. We proceed to revise Peperkamp et al.'s framework and circumvent these limitations, reducing the task of the acquisition of phonemes to a partitioning-clustering problem.

Let us emphasize immediately, however, that no experiment reported in this dissertation is to be considered as a complete and plausible model of early language acquisition. The reason for this is twofold. First, there is no guarantee that the algorithms and data structures used in our simulations bear any resemblance to the cognitive processes and mental representations available to or used by infants. Second, we focused on providing the first empirical bounds on the learnability of allophony in Peperkamp et al.'s framework—relying, for instance, on supervised learning techniques and non-trivial simplifying assumptions regarding the nature of phonological processes. Though motivated by psycholinguistic considerations, the present study is thus to be considered a contribution to data-intensive experimental linguistics.

1.3 Structure of the dissertation

The body of this dissertation is divided into six main chapters, delimited according to the different data representations and classification or clustering tasks we examined.

In Chapter 2, we introduce the major concepts at play in the present study—viz. those of phone, phoneme, and allophone. We also review in this chapter the state of the art in the field of computational modeling of early phonological acquisition. Chapter 3 is an introduction to the corpus of Japanese speech we used throughout this study. Here, the focus is on discussing our preprocessing choices—especially regarding the definition of the phonemic inventory, the mechanisms we used to control the limits of the phonetic similarity between the allophones of a given phoneme, and how our data eventually relate to theoretical descriptions of the phonology of Japanese. In Chapter 4, we report our preliminary experiments on the acquisition of allophony. We first define the core concepts of our contribution—viz. empirical measures of dissimilarity between phones referred to as indicators of allophony. Then, we evaluate these indicators in experiments similar to the ones carried out by Peperkamp et al. (2006; and subsequent studies), and try to predict whether two phones are realizations of the same phoneme. Chapter 5 is divided into two main sections. In the first section, we discuss the limitations of Peperkamp et al.'s pairwise framework, as well as our arguments in favor of a transition toward the fundamental proposition of this study, i.e. not only predicting whether two phones are realizations of the same phoneme but, if so, of which phoneme they both are realizations. In the section that follows, we report various transitional experiments where, using the very same data as in Chapter 4, we attempt to classify phone pairs into phoneme-like categories. In Chapter 6, we start with a formal description of the techniques we used to obtain a novel, pair-free representation for the data at hand that is more suitable to the prediction of phonemehood. We then go on to report various classification and clustering experiments aiming at partitioning a set of phones into a putative phonemic inventory. Finally, Chapter 7 contains a general conclusion in which we discuss the contributions of the present study, as well as the limitations and possible improvements to our computational model of the early acquisition of phonemes.

CHAPTER 2
OF PHONES AND PHONEMES

When embarking on a study of sounds and sound systems, it is important to first define the objects with which we will be dealing. As we are dealing with both tangible sound objects and abstract sound categories, we need to define just what these objects are, as well as how we conceive the notion of sound category itself. Therefore, the aim of this chapter is to introduce the major concepts at play in the present study (namely those of phone, phoneme, and allophone), and to review the state of the art in the area of computational modeling of early phonological acquisition.

2.1 The sounds of language

The subfields of linguistics can ideally be organized as a hierarchy wherein each level uses the units of the lower level to make up its own, ranging from the tangible yet meaningless (at least when considered in isolation) acoustic bits and pieces studied in phonetics, to the meaningful yet intangible abstract constructs studied in semantics and discourse. In this global hierarchy where discourses are made of utterances, utterances are made of words, etc., phonology and phonetics can be thought of as the two (tightly related) subfields concerned with the smallest units: sounds. Before giving precise definitions for the sounds of language, it is worth emphasizing that, as far as phonology and phonetics are concerned, words are made of sounds (more precisely, phonemes) rather than letters (graphemes). Not all natural languages have written forms (Eifring & Theil, 2004) and, if they do, the writing system and the orthography of a given language are nothing but agreed-upon symbols and conventions that do not necessarily reflect any aspect of the underlying structure of that language.

2.1.1 Phones and phonemes

In this study, we are interested in the tangible, vocalized aspects of natural languages. In the context of verbal communication, the speaker produces an acoustic signal with a certain meaning and, if communication is successful, the hearer is able to retrieve the intended meaning from the received signal. Communication is possible if, among other things, the hearer and the speaker share the common knowledge of the two-way coding scheme that is (unconsciously) used to transform the message into sound, and vice versa. Such a process is one among many examples of Shannon's (1948) noisy channel model which describes how communication can be successfully achieved when the message is contaminated to a certain extent by noise or variation to a norm. Indeed, as mentioned by Coleman (1998; p. 49), it was remarked early on by linguists that

> ❝the meaning of a word or phrase is not signalled by exactly *how* it is pronounced —if it was, physiological differences between people and the acoustic consequences of those differences would make speech communication almost impossible— but how it *differs* from the other words or phrases which might have occurred instead.❞

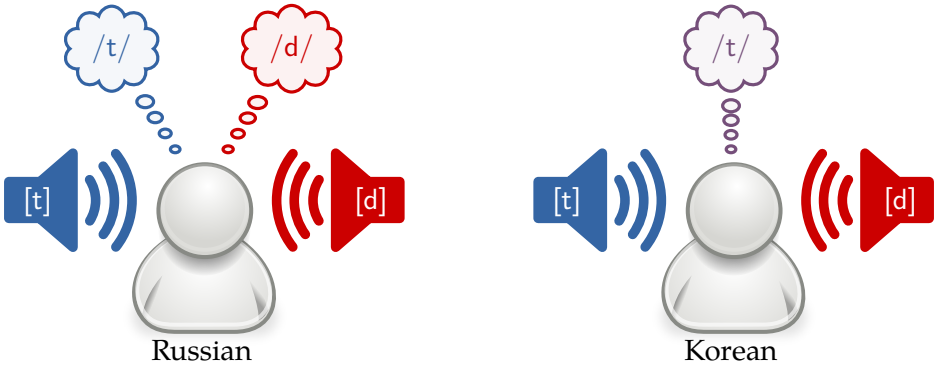


Figure 2.1 — Mental representation of the phones [t] and [d] for speakers of Russian and speakers of Korean, after Kazanina et al. (2006).

Two important aspects of natural languages are highlighted in this quotation. On one hand, the quote pinpoints the fact that linguistic units are not relevant or useful by themselves, but as the commutable components of a system. On the other hand, it emphasizes the need for a dichotomy between the exact acoustic realization of the message and a more abstract representation. In other words, not all acoustic information is relevant linguistic information. The concepts of phone and phoneme embody this dichotomy: whereas *phonemes* are abstract sound categories, *phones*are tangible realizations of the phonemes.

Phonemes Phonology and sound systems have been studied under various (sometimes competing) frameworks and theories, each of them promoting its own representations, rules, principles, or constraints (e.g. Chomsky & Halle, 1968; Dell, 1985; Coleman, 1998; Kager, 1999; Goldsmith et al., 2011). However, it seems fair to assume that, no matter the framework, phonemes are among the most elementary linguistic units manipulated by phonologists. Indeed, the International Phonetic Association (1999) defines the phoneme as

> ❝the smallest segmental unit of sound employed to form meaningful contrasts between utterances.❞

Minimal pairs are the typical example of such meaningful contrasts: in English, for example, word pairs such as /kæt/ ∼ /bæt/ (*cat* and *bat*), /kæt/ ∼ /kɪt/ (*cat* and *kit*), or /sɪŋk/ ∼ /θɪŋk/ (*sink* and *think*) can be distinguished thanks to the contrasts between the phoneme pairs /k/ ∼ /b/, /æ/ ∼ /ɪ/, and /s/ ∼ /θ/, respectively. Put another way, not distinguishing one phoneme from another, for example /s/ ∼ /θ/ as some European-French- or Chinese-speaking learners of English do (Brannen, 2002; Rau et al., 2009), would inevitably yield to the confusion of words contrasting on these phonemes.

If native speakers of a given language may not be able to distinguish some phonemes of another language, it is because each language's phonology is organized around its own finite (and somewhat small) set of phonemes, a.k.a. the language's phonemic inventory. Phonemic inventories vary significantly across languages, both in size and in content (Tambovtsev & Martindale, 2007; Atkinson, 2011; Wang et al., 2012), and distinctive contrasts in one language may well not be distinctive in another. Evidence of such contrasts and of the psychological reality of phonemes is presented by Kazanina et al. (2006) who demonstrated that early auditory brain responses are shaped by the functional role of the sounds in the listener's language, as pictured in Figure 2.1. Using magnetoencephalographic brain recordings, Kazanina et al. showed that hearing the very same phones [t] and [d] activates distinct response patterns for speakers of Russian and speakers of Korean: whereas the brain recordings of the speakers of Russian showed divergent responses for [t] and [d], those of speakers of Korean showed no significantly divergent responses. In the same study, speakers of Korean also showed great difficulty in consciously discriminating [t] and [d]. This discrepancy in performance is due to the fact that the stimuli [t] and [d] are realizations of two distinct phonemes /t/ and /d/ in Russian, whereas they are both

realizations of a single phoneme in Korean, here written as /t/. Indeed, as emphasized by Tobin (1997; p. 314):

> ❝native speakers are clearly aware of the *phonemes* of their language but are both unaware of and even shocked by the plethora of *allophones* and the minutiae needed to distinguish between them.❞

We previously stated that speakers of English are able to distinguish word pairs such as /kæt/ ∼ /bæt/ because of the contrast between the phonemes /k/ ∼ /b/. In fact, the reciprocal of this statement may yield a more accurate definition of phonemes as linguistic units: in English, the contrast between sound categories such as /k/ ∼ /b/ is said to be phonemic because it allows the speakers to distinguish word pairs such as /kæt/ ∼ /bæt/. Phonemes do not exist for their own sake, but for the purpose of producing distinct forms for messages with distinct meanings.

Phones Whereas each language's sound system comprises a limited number of phonemes, the number of phones (a.k.a. speech sounds or segments) is unbounded. Due to many linguistic and extra-linguistic factors such as the speaker's age, gender, social background, or mood, no two realizations of the same abstract word or utterance can be strictly identical. Moreover, and notwithstanding the difficulty of setting phone boundaries (Kuhl, 2004), we define phones as nothing but phoneme-sized chunks of acoustic signal. Thence, as argued by Lyons (1968; p. 100), the concept of phone accounts for a virtually infinite collection of language-independent objects:

> ❝The point at which the phonetician stops distinguishing different speech sounds is dictated either by the limits of his own capacities and those of his instruments or (more usually) by the particular purpose of the analysis.❞

To draw an analogy with computer science and information theory, phonemes can be thought of as the optimal lossless compressed representation for human speech. Lossless data compression refers to the process of reducing the consumption of resources without losing information, identifying and eliminating redundancy (Wade, 1994; p. 34). In the case of verbal communication, the dispensable redundancy is made of the fine-grained acoustic information that is responsible for all conceivable phonetic contrasts, while the true information is made of the language's phonemic contrasts. Whereas eliminating phonetic contrasts, e.g. [kæt] ∼ [kɑt] or [kæt] ∼ [k̰æt], would not result in any loss of linguistic information, no further simplification could be applied once the phonemic level has been reached without risking to confuse words, e.g. /kæt/ ∼ /bæt/. The concept of phoneme might be thus defined as the psychological representation for an aggregate of confusable phones. This conception of phonemes as composite constructs has seldom been emphasized in the literature, though a notable example is Miller (1967; p. 229) who, describing the phonemic inventory of Japanese, denotes as "/i/ the syllabic high front *vowels*" (emphasis added). The point at which we stop distinguishing different phones in the present study will be presented and discussed in Chapter 3.

Although any phoneme can virtually be realized by an infinite number of distinct phones, the one-to-many relation between sound categories and sound tokens is far from being random or arbitrary. First and foremost, the realizations of a given phoneme are phonetically (or acoustically, we consider both terms to be synonyms) similar, to a certain extent. However, it is worth noting that the applicability of phonetic similarity as a criterion for the definition of phonemehood has been sorely criticized by theoretical linguists such as Austin (1957; p. 538):

> ❝Phonemes are phonemes because of their function, their distribution, not because of their phonetic similarity. Most linguists are arbitrary and ad hoc about the physical limits of phonetic similarity.❞

On the contrary, we argue in favor of the relevance of phonetic similarity on the grounds that it should be considered an observable consequence rather than an underlying cause of phonemehood. We suggest the following reformulation of the phonetic criterion, already hinted at by Machata & Jelaska (2006): more than the phonetic similarity between the realizations of a given phoneme, it is the phonetic dissimilarity between the realizations of different phonemes that may be used as a criterion for the definition of phonemehood. Phonemes are phonemes because

of their function, and that function can only be carried out effectively if their realizations can be distinguished without ambiguity. This conception of the relation between phonetic reality and phonological abstraction is reminiscent of the compactness hypothesis (Duin, 1999; Pękalska et al., 2003; and references therein) underlying most if not all classification and clustering problems. Broadly speaking, the compactness hypothesis can be formulated as follows: similar objects have to be close in their representation space. Assuming so, and in the perspective of classifying the objects, objects whose representations are close enough would be assigned to the same class. Thus, if some objects whose representations are close enough do not belong to the same class, then their classes most certainly overlap in this representation and cannot be distinguished. We think this reasoning holds in the case at hand: in any language, phones of different phonemes have to be sufficiently dissimilar so that hearers can recover the underlying phonemic inventory of the language without uncertainty. As we will argue in Section 2.2.2 and Chapter 4, phonetic similarity is not a sufficient criterion for the definition of the phonemic inventory of a language. Nevertheless, we consider counterintuitive any approach to phonological studies that choses to dismiss outright the acoustic form of language.

2.1.2 Allophony

We previously stated that the realizations of a given phoneme are phonetically similar, but only to a certain extent. Indeed, in any language, the realizations of a given phoneme are constrained by various physiological and phonological phenomena. Broadly speaking, the realization of a given phoneme in a given word or utterance is shaped and constrained by (the realizations of) the surrounding phonemes.

Soundalikes Assimilation and coarticulation are two linguistic concepts that refer to the contextual variability of phonemes' realizations. Although some scholars have argued that a strong dichotomy between assimilation and coarticulation should be made (e.g. Chomsky & Halle, 1968), it is not clear whether these terms refer to distinct processes, or to different points of view on the same process. Farnetani (1997; pp. 371 and 376), for example, gives the following definitions:

> ❝During speech the movements of different articulators for the production of successive phonetic segments overlap in time and interact with one another. As a consequence, the vocal tract configuration at any point in time is influenced by more than one segment. This is what the term "coarticulation" describes. [...] Coarticulation may or may not be audible in terms of modifications of the phonetic quality of a segment. [...] Assimilation refers to contextual variability of speech sounds, by which one or more of their phonetic properties are modified and become similar to those of the adjacent segments.❞

Confronting so much indeterminacy, we reiterate our statement (Boruta, 2011b) that coarticulation and assimilation are two poles on a continuum, ranging from accidental and barely perceptible sound changes of interest mainly to phoneticians (coarticulation) to systematic and substantial sound changes of interest mainly to phonologists (assimilation).

Phonological rules In the context of writing phonological grammars, assimilation processes are often described as rules expressing how a given phoneme is realized as a given phone in a given context. For instance, consider the following description of the possible realizations of the consonant /t/ in English (Lyons, 1968; Jurafksy & Martin, 2009), simplified for the sake of the example. Before a dental consonant, the consonant /t/ is realized as a dentalized [t̪] sound and, for example, the word *eighth* is pronounced [eɪt̪θ]. Between two vowels, this consonant is realized as an alveolar flap [ɾ] sound and, for example, the word *kitten* is pronounced [kɪɾən]. As the initial phoneme of a word, the consonant is realized as an aspirated [tʰ] sound and, for example, the word *tunafish* is pronounced [tʰjuːnəfɪʃ]. In all other contexts, the consonant /t/ is realized as a plain [t] sound and, for example, the word *starfish* is pronounced [stɑːfɪʃ]. It is

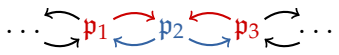


Figure 2.2 — Schematic representation of the contextual variability for three phonemes $\mathfrak{p}_1$, $\mathfrak{p}_2$, and $\mathfrak{p}_3$. An arrow from one phoneme to another indicates that the former may influence the realization of the latter. While phonological descriptions often highlight the realization of a given phoneme in a given context (here in red), allophony and coarticulation actually alter the whole sequence: each phoneme is an attested context of its attested contexts (here in blue).

worth noting that, even from this simple example, one can already observe the manifold nature of phonological contexts: some consist of either the preceding or the following phoneme, some consist of both the preceding and the following phonemes, and some consist of yet-unmentioned abstract phonological objects such as word boundaries.

In the present study, we are interested in the relation that brings together all the realizations of a given phoneme, for example [t], [t̪], [tʰ], and [ɾ] in the case of /t/ in English. The possible phones of a given phoneme are usually referred to as *allophones* (e.g. Lyons, 1968), while the relation between these phones is referred to as *allophony*. As will be discussed in greater detail later in this study, we are interested in learning the phonemic inventory of a language—that is to say that we want to be able to predict whether two phones are allophones of a phoneme in the language at hand and, in that case, to determine of which phoneme they both are realizations. Although such tasks may at first seem trivial because of the limited number of classes (i.e. the phonemes) and the inherent similarity between the members of a given class (the allophones), it is worth noting that the apparent conciseness of theoretical phonological descriptions is due to the fact that allophonic grammars are often presented in the context of speech production, rather than speech perception. In such a top-down (if not generative) context, one can readily use phonemes and even phoneme classes to describe the realizations of other phonemes—stating for example, as we did in the previous paragraph, that /t/ is realized as [ɾ] between two vowels in English. Such a description is only possible if the whole phonemic inventory is known: indeed, not only do we need to know that [t], [t̪], [tʰ], and [ɾ] collectively make for what is denoted as /t/, but also that [ɾ] is only attested between the realizations of other phonemes collectively known as vowels, which are themselves defined in contrast with consonants such as /t/. This apparent problem of circularity may be more clearly examplified by the simple observation that each phoneme is an attested context of its attested contexts, as presented in Figure 2.2. In other words, each phoneme is at the same time the target of an allophonic rule, and (part of) the application context of allophonic rules targeting its contexts. For instance in the English word /kɪtən/, /t/ is both the target of an allophonic rule whose context is made of /ɪ/ and /ə/, and part of the contexts of two allophonic rules whose respective targets are /ɪ/ and /ə/.

Assuming the position of a naive learner, how can one break these reciprocal patterns of contextual variability and bootstrap the phonemic inventory of the language at hand? In particular, how do infants learn the phonemes of their native language, and how is this aspect of early language acquisition related to the discovery of word forms and word meanings? These higher-level questions motivate the present study. In the next section, we will review related work and discuss in greater detail our study's departures from previously published studies, as well as our methods and assumptions.

2.2 Early phonological acquisition: state of the art

As outlined in the previous section, the two-way mapping between tangible phones and abstract phonemes is complex. Furthermore, being able to distinguish between phonemes seems to be a necessary skill for word recognition and, as a consequence, the acquisition of phonological knowledge appears to be one of the earliest stages of language acquisition. How much variability do infants have to tackle to learn the phonemic inventory of their native language? How, for example, do English-learning infants learn that [kæt] and [kat] refer to the same animal species,

whereas [kæt] and [bæt] do not? What kind of cues may they rely on to learn that [sɪŋkɪŋ] and [θɪŋkɪŋ] can not refer to the same action?

Broadly speaking, research on early language acquisition falls into one of two categories: behavioral experiments and computational experiments. On one hand, behavioral experiments are used to assess infants' response to linguistic stimuli; however, in order to ensure that observed responses are due solely to the specific linguistic phenomenon studied in the experiment, stimuli have to be thoroughly controlled, and experimental material often consists of synthesized speech or artificial languages. On the other hand, computational experiments allow for the use of unaltered (yet digitized) natural language. However, conclusions regarding language acquisition (or any other psychological process) drawn from the output of a computer program should be considered with all proper reservations, as there is no guarantee that the algorithms and data structures used in such simulations bear any resemblance to the cognitive processes and mental representations available to or used by infants. Because of this trade-off between the ecological validity of the learner and that of the learning material, in vivo and in silico experiments are complementary approaches for research on language acquisition. The outline of this section follows this dichotomy: we will first provide a brief review of behavioral results in order to set the scene, and we will then review computational results in order to discuss how our study departs from previously reported research efforts.

Babies, children and infants Throughout this study, young human learners will be uniformly referred to as *infants*, mainly because of the word's latin etymology, *infans*, meaning "not able to speak." We henceforth use *infant-directed speech* as an umbrella term for *motherese*, *child-directed speech*, or any other word used to describe infants' linguistic input.

It is also worth emphasizing immediately that, for the sake of simplicity, the present study focuses on early language acquisition in a monolingual environment. As interesting as bi- or plurilingualism may be, learning the phonemic inventory of a single language is a daunting enough task, as we will explain in the rest of this study.

2.2.1 Behavioral experiments

Although an extensive review of the literature on early language acquisition would require a monograph of its own, we will here review the major results in the field, focusing on speech perception and the emergence of phonemes. Furthermore, and for the sake of simplicity, we will mainly give references to a small number of systematic reviews (Plunkett, 1997; Werker & Tees, 1999; Peperkamp, 2003; Clark, 2004; Kuhl, 2004), rather than to all original studies. For a review focusing on the acquisition of speech production, see Macneilage (1997).

Cracking the speech code Remarkable results about infants' innate linguistic skills have highlighted their ability to discriminate among virtually all phones, including those illustrating contrasts between phonemes of languages they have never heard (Kuhl, 2004). Nevertheless, some adaptation to their native language's phonology occurs either in utero or immediately after birth; indeed, it has been observed that infants aged 2 days show a preference for listening to their native language (Werker & Tees, 1999). Afterwards, exposure to a specific language sharpens infants' perception of the boundaries between the phonemic categories of that language. For example, English-learning infants aged 2–4 months are able to perceptually group realizations of the vowel /i/ uttered by a man, woman, or child in different intonative contexts, distinguishing these phones from realizations of the vowel /a/ (Werker & Tees, 1999). As they improve their knowledge of their native language's phonology, infants also gradually lose the ability to discriminate among all phones. While it has been shown, for instance, that English-learning infants aged 6–8 months are able to distinguish the glottalized velar vs. uvular stop contrast [kˈ] ∼ [qˈ] in Nthlakapmx, a Pacific Northwest language, this ability declines after the age of 10–12 months (Plunkett, 1997). It is also worth noting that, throughout the course of language acquisition,

infants learn not only the phonemes of their native language, but also how to recognize these phonemes when uttered by different speakers and in different contexts (Kuhl, 2004), a task that has proven to be troublesome for computers (e.g. Woodland, 2001). As a whole, learning phonemes can be thought of as learning not to pay attention to (some) detail.

Stages of interest in Kuhl's (2004) universal timeline of speech perception development can be summarized as follows. From birth to the age of 5 months, infants virtually discriminate the phonetic contrasts of all natural languages. Then, language-specific perception begins to take place for vowels from the age of 6 months onwards. This is also the time when statistical learning can first be observed, most notably from distributional frequencies. The age of 8 months marks a strong shift in the transition from universal to language-specific speech perception: now also relying on transitional probabilities, infants are able to recognize the prosody (viz. the typical stress patterns) and the *phonotactics* (i.e. the admissible sound combinations) of their native language. From the age of 11 months onwards, infants' sensitivity to the consonants of their native language increases, at the expense of the (universal) sensitivity to the consonants in foreign languages. Finally, infants master the major aspects of the grammar of their native language by the age of 3 years (Peperkamp, 2003).

Statistical learning Infants' ability to detect patterns and regularities in speech is often referred to as statistical learning, i.e. the acquisition of knowledge through the computation of information about the distributional frequency of certain items, or probabilistic information in sequences of such items (Kuhl, 2004). In particular under a frequentist interpretation of probabilities (Hájek, 2012), this appealing learning hypothesis is corroborated by various experiments where language processing has been showed to be, in adults as in infants, intimately tuned to frequency effects (Ellis, 2002). Virtually all recent studies on early language acquisition whose models or assumptions rely on statistical learning build upon the seminal studies by Saffran et al. who showed that 8-month-old infants were able to segment words from fluent speech making only use of the statistical relations between neighboring speech sounds, and after only two minutes of exposure (Saffran et al., 1996; Aslin et al., 1998; Saffran, 2002). However, as emphasized by Kuhl (2004; pp. 831–832), infants' learning abilities are also highly constrained:

> ❝Infants can not perceive all physical differences in speech sounds, and are not computational slaves to learning all possible stochastic patterns in language input. [...] Infants do not discriminate all physically equal acoustic differences; they show heightened sensitivity to those that are important for language.❞

Although it would be beyond the stated scope of this research to enter into the debate of nature vs. nurture, it is worth noting that various authors have proposed that the development of speech perception should be viewed as an innately guided learning process (Jusczyk & Bertoncini, 1988; Gildea & Jurafsky, 1996; Plunkett, 1997; Bloom, 2010). Bloom, for instance, highlights this apparent predisposition of infants for learning natural languages as follows:

> ❝One lesson from the study of artificial intelligence (and from cognitive science more generally) is that an empty head learns nothing: a system that is capable of rapidly absorbing information needs to have some prewired understanding of what to pay attention to and what generalizations to make. Babies might start off smart, then, because it enables them to get smarter.❞

Furthermore, as prominent as Saffran et al.'s (1996) results may be, they should not be overinterpreted. Although it is a necessity for experimental studies to use controlled stimuli, it is worth pointing out that the stimuli used by Saffran et al. consisted in four three-syllable nonsense words (viz. /tupiro/, /bidaku/, /padoti/, and /golabu/ for one of the two counterbalanced conditions) whose acoustic forms were generated by a speech synthesizer in a monotone female voice at a constant rate of 270 syllables per minute. As stated by the authors, the only cue to word boundaries in this artificial language sample were the transitional probabilities between all syllable pairs. By contrast, natural infant-directed speech contains various cues to different aspects of the language's grammar: stress, for example, is one of many cues for the discovery of word

boundaries in various languages, and it has been shown to collide with statistical regularities in (more plausible) settings where infants have to integrate multiple cues (Johnson & Jusczyk, 2001; Thiessen & Saffran, 2003). Additionally, not all statistically significant co-occurrence patterns are relevant cues for learning the grammar of either the language at hand or, possibly, any other language. Indeed, as emphasized by Gambell & Yang (2004; p. 50):

> ❝While infants seem to keep track of statistical information, any conclusion drawn from such findings must presuppose children knowing what kind of statistical information to keep track of. After all, an infinite range of statistical correlations exists in the acoustic input: e.g., What is the probability of a syllable rhyming with the next? What is the probability of two adjacent vowels being both nasal?❞

In the present study, we endorse the aforementioned views that the development of speech perception is an innately guided learning process and that statistical learning is a key component of early language acquisition. The following section reviews previously published modeling efforts with the aim of making each study's major assumptions explicit, thus specifying how this study departs from the state of the art.

2.2.2 Computational experiments

As noted by Peperkamp (2003), experimental research on speech perception has shown a strong focus on the acquisition and the mental representation of phonemes and allophones. Unfortunately, this observation does not hold for computational studies. Indeed, while a limited number of studies attempting to model how infants may learn the phonemes of their native language have been reported so far, a tremendous number of studies on the acquisition of word segmentation strategies have been published (e.g. Olivier, 1968; Elman, 1990; Brent & Cartwright, 1996; Christiansen et al., 1998; Brent, 1999; Venkataraman, 2001; Johnson, 2008b; Goldwater et al., 2009; Pearl et al., 2010; to cite but a few). In our opinion, this ongoing abundance of computational models and experiments on word segmentation is due to the fact that merely splitting sequences of characters requires more skills in stringology than in psychology or linguistics. Indeed, until very recently (Rytting et al., 2010; Daland & Pierrehumbert, 2011; Boruta et al., 2011; Elsner et al., 2012), most—if not all—models of word segmentation have used idealized input consisting in phonemic transcriptions and, as we have argued in a previous study (Boruta et al., 2011; p. 1):

> ❝these experiments [...] make the implicit simplifying assumption that, when children learn to segment speech into words, they have already learned phonological rules and know how to reduce the inherent variability in speech to a finite (and rather small) number of abstract categories: the phonemes.❞

The ultimate goal of the project reported in the present study is to develop and validate a computational model of the acquisition of phonological knowledge that would precisely account for this assumption, mapping each phone to the phoneme of which it is a realization.

Before reviewing related work, it is worth emphasizing immediately that the present study departs from most research efforts on allophony and sound variability on one major point: we are interested in gaining valuable insights on human language and human language acquisition. Although our results were obtained through computational experiments, we have for this reason little interest for performance-driven methods for phoneme recognition, as developed and used in the field of automatic speech recognition (e.g. Waibel et al., 1989; Matsuda et al., 2000; Schwarz et al., 2004; Huijbregts et al., 2011). By contrast, our subject matter is language and, thence, the present study is to be considered a contribution to data-intensive experimental linguistics, in the sense of Abney (2011).

Related work Research efforts on models of the acquisition of phonology have been rather scarce and uncoordinated: to our knowledge, no shared task or evaluation campaign has ever been organized on such topics, while different studies have seldom used the same data or algorithms. We will briefly review previously published models in this section, highlighting the

similarities and discrepancies between them and the present study. For general discussions of category learning and phonological acquisition, see Boersma (2010) and Ramus et al. (2010).

At the pinnacle of structuralism, Harris (1951, 1955) presented a procedure for the discovery of phonemes (and other linguistic units) that relies on successor counts, i.e. on the distributional properties of phones within words or utterances in a given language. Unfortunately, this discovery procedure is only meant to be applied by hand as too many details (e.g. concerning data representation or the limit of phonetic similarity) were left unspecified as to be able to implement and validate a full computational model.

To our knowledge, the first complete algorithms capable of learning allophonic rules were proposed by Johnson (1984), Kaplan & Kay (1994) and Gildea & Jurafsky (1996): all three algorithms examine the machine learning of symbolic phonological rules à la Chomsky & Halle (1968). However, as noted by Gildea & Jurafsky, these algorithms include no mechanism to account for the noise and the non-determinism inherent to linguistic data. An additional limitation of Gildea & Jurafsky's algorithm is that it performs supervised learning, i.e. it requires a preliminary training step during which correct pairs of phonetic and phonological forms are processed; yet, performing unsupervised learning is an essential plausibility criterion for computational models of early language acquisition (Brent, 1999; Alishahi, 2011; Boruta et al., 2011) as infant-directed speech does not contain such ideal labeled data.

A relatively high number of models of the acquisition of phonological knowledge (Tesar & Smolensky, 1996; Boersma & Levelt, 2000, 2003; Boersma et al., 2003; Hayes, 2004; Boersma, 2011; Magri, 2012) have been developed in the framework of optimality theory (Prince & Smolensky, 1993; Kager, 1999). Tesar & Smolensky's (1996) model, for instance, is able to learn both the mapping from phonetic to phonological forms and the phonotactics of the language at hand. It is however worth mentioning that optimality-theoretic models make various non-trivial assumptions—viz. regarding the availability of a phonemic inventory, distinctive features, word segmentation, and a mechanism recovering underlying word forms from their surface forms. For this reason, such models are not comparable with the work presented in this dissertation, as we specifically address the question of the acquisition of a given language's phonemic inventory.

In an original study, Goldsmith & Xanthos (2009) addressed the acquisition of higher-level phonological units and phenomena, namely the pervasive vowel vs. consonant distinction, vowel harmony systems, and syllable structure. Doing so, they assume that the phonemic inventory and a phonemic representation of the language at hand are readily available to the learner. Goldsmith & Xanthos' matter of concern is hence one step ahead of ours.

Building upon a prior theoretical discussion that infants may be able to undo (some) phonological variation without having acquired a lexicon (Peperkamp & Dupoux, 2002), Peperkamp et al. have developed across various studies a bottom-up, statistical model of the acquisition of phonemes and allophonic rules relying on distributional and, more recently, acoustic and proto-lexical cues (Peperkamp et al., 2006; Le Calvez, 2007; Le Calvez et al., 2007; Dautriche, 2009; Martin et al., 2009; Boruta, 2009, 2011b). Revamping Harris' (1951) examination of every phone's attested contexts, Peperkamp et al.'s (2006) distributional learner looks for pairs of phones with near-complementary distributions in a corpus. Because complementary distributions is a necessary but not sufficient criterion, the algorithm then applies linguistic constraints to decide whether or not two phones are realizations of the same phoneme, checking, for example, that potential allophones share subsegmental phonetic or phonological features (cf. Gildea & Jurafsky's faithfulness constraint). Further developments of this model include replacing a priori faithfulness constraints by an empirical measure of acoustic similarity between phones (Dautriche, 2009), and examining infants' emerging lexicon (Martin et al., 2009; Boruta, 2009, 2011b). Consider, for example, the allophonic rule of voicing in Mexican Spanish, as presented by Le Calvez et al. (2007), by which /s/ is realized as [z] before voiced consonants (e.g. *feliz Navidad*, "happy Christmas," is pronounced as [feliz͡nabidad], ignoring other phonological processes for the sake of the example) and as [s] elsewhere (e.g. *feliz cumpleaños*, "happy birthday," is pronounced as [felis͡kumpleaɲos]). Although it introduces the need for an ancillary word

segmentation procedure, Martin et al. (2009) showed that tracking alternations on the first or last segment of otherwise identical word forms such as [feliz] ∼ [felis] (which are not minimal pairs stricto sensu) is a relevant cue in order to learn allophonic rules.

Except for Dautriche's (2009) extension of Peperkamp et al.'s model, a stark limitation of all aforementioned models is that they operate on transcriptions, and not on sounds. As we argued in a previous study (Boruta, 2011b), phones are nothing but abstract symbols in such approaches to speech processing, and the task is as hard for [a] ∼ [ḁ] as it is for [ɥ] ∼ [k]. Despite the fact that formal language theory has been at the heart of computational linguistics and, especially, computational phonology (Kaplan & Kay, 1994; Jurafksy & Martin, 2009; Wintner, 2010; and references therein), we reiterate our argument that legitimate models of the acquisition of phonological knowledge should not disregard the acoustic form of language.

Several recent studies have addressed the issue of learning phonological knowledge from speech: while Yu (2010) presented a case study with lexical tones in Cantonese, Vallabha et al. (2007) and Dillon et al. (2012) focused on learning phonemic categories from English or Japanese infant-directed speech and adult-directed Inuktitut, respectively. However, in all three studies, the linguistic material used as the models' input was hand-annotated by trained phoneticians, thus hindering the reproducibility of these studies. Furthermore, and even if a model can only capture certain aspects of the process of language acquisition, highlighting some while muting others (Yu, 2010; p. 11), both studies attempting to model the acquisition of phonemes restricted the task to the acquisition of the vowels of the languages at hand (actually a subset of the vowel inventories in Vallabha et al.'s study). Although experimental studies have suggested that language-specific vowel categories might emerge earlier than other phonemes (Kuhl, 2004; and references therein), we consider that the phonemic inventory of any natural language is a cohesive and integrative system that can not be split up without reserve. These computational studies offer interesting proof of concepts, but they give no guarantee as to the performance of their respective models in the (plausible) case in which the learner's goal is to discover the whole phonemic inventory.

The overall goal of the present study is to build upon Peperkamp et al.'s experiments to propose a model of the acquisition of phonemes that supplements their distributional learner (Peperkamp et al., 2006; Le Calvez, 2007; Le Calvez et al., 2007; Martin et al., 2009; Boruta, 2009, 2011b) with an examination of available acoustic information (Vallabha et al., 2007; Dautriche, 2009; Dillon et al., 2012). To do so, we introduce a (somewhat artificial) dichotomy between the concepts of allophony and phonemehood, mainly because of the discrepancies between the state of the art and the aim of this project: whereas Peperkamp et al. have focused on predicting whether two phones are realizations of the same phoneme (what we will refer to as allophony), we are eventually interested in predicting which phoneme they both are realizations of (what we will refer to as phonemehood). Bridging, to some extent, the gap between symbolic (i.e. phonology-driven) and numeric (i.e. phonetics-driven) approaches to the acquisition of phonology, we will also address the issue of the limits of phonetic similarity by evaluating our model with data undergoing different degrees of phonetic and phonological variability.

Models of the human mind Cognitive modeling and so called psychocomputational models of language acquisition have received increasing attention in the last decade, as shown by the emergence (and continuance) of specialized conferences and workshops such as *Cognitive Modeling and Computational Linguistics* (CMCL), *Psychocomputational Models of Human Language Acquisition* (PsychoCompLA), or *Architectures and Mechanisms for Language Processing* (AMLaP). Because of the diversity in linguistic phenomena, theories of language acquisition, and available modeling techniques, developing computational models of early language acquisition is an intrinsically interdisciplinary task. Though computational approaches to language have sometimes been criticized by linguists (e.g. Berdicevskis & Piperski, 2011), we nonetheless support Abney's (2011) argumentation that data-intensive experimental linguistics is genuine linguistics, and we will hereby discuss the major points that distinguish a computational model of language processing

from a linguistic theory.

The objection that our conclusions are merely drawn from the output of a computer program (rather than from observed behavioral responses) was raised various times during early presentations of the work reported in this dissertation. Although this is true, by definition, of any computational study, it is not a strong argument, as pointed out by Hodges (2007; pp. 255–256; see also the discussion by Norris, 2005):

> ❝The fact that a prediction is run on a computer is not of primary significance. The validity of the model comes from the correctness of its mathematics and physics. Making predictions is the business of science, and computers simply extend the scope of human mental faculties for doing that business. [...] Conversely, if a theory is wrong then running it on a computer won't make it come right.❞

We nevertheless concede that a major issue faced by computational linguists is that our linguistic material is not the kind of data that computers manipulate natively: while, broadly speaking, linguists are interested in sounds and words, computers only manipulate zeros and ones. Hence, to the bare minimum, the speech stream needs to be digitized; phones are represented as vectors, and words as sequences of such vectors. Put another way, words fly away, while numbers remain. Depending on the chosen numerical representation, adapting or transforming the input is not without consequences, as noted by Pękalska & Duin (2005; p. 163):

> ❝Such a simplification of an object to its numerical description (i.e. without any structural information) precludes any inverse mapping to be able to retrieve the object itself.❞

Therefore, throughout this study, a special emphasis will be put on discussing how phones and phonemes are represented in our simulations, as well as how these representations may impact the performance of the models.

In addition to data representation, working hypotheses and modeling assumptions are another pitfall for computational linguists. Indeed, an algorithm may impose preconditions on the input data (e.g. requirements for statistical independence of the observations, a given probability distribution, or a minimum number of observations). Different algorithms solving the same problem may impose different preconditions on the data and, if one or more preconditions are violated (or arbitrarily waived), the behavior of an algorithm may become flawed or unspecified. A potential issue lies in the fact that not all preconditions can be verified automatically: binomial logistic regression, for example, assumes that the observations are independent (i.e. the outcome for one observation does not affect the outcome for another; Agresti, 2007; p. 4), but this property can not be verified without expert knowledge, and any implementation would be able to fit a binomial logistic regression model to some observations, no matter whether they are truly independent or not. In other words, the correctness of the program does not guarantee the correctness of the experiment, thus illustrating the previous quotation from Hodges (2007). Furthermore, making assumptions explicit is a methodological advantage of computational models over theories, as argued by Alishahi (2011; p. 5):

> ❝This property distinguishes a computational model from a linguistic theory, which normally deals with higher-level routines and does not delve into details, a fact that makes such theories hard to evaluate.❞

This discrepancy may be reformulated in terms of Marr's (1982) levels of analysis. Whereas linguistic theories may only probe the computational level (i.e. a description of the problem and of the global logic of the strategy used to solve it), computational models must provide a thorough and extensive discussion of the algorithmic level (i.e. a description of the operated transformations and of the representations for the input and the output). The last level, viz. the implementation level, is beyond the stated scope of this research as it describes how the system is physically realized and thus, in the case at hand, belongs to the field of neurolinguistics.

For the aforementioned reasons, a special emphasis will also be put on making modeling assumptions explicit throughout this study, highlighting them in the way that proofs or theorems are often highlighted in mathematics. For instance, we have already made the following assumptions through the course of this chapter:

Assumption 2.1 The development of speech perception by infants is neither fully innate nor fully acquired, but an innately guided learning process (cf. Jusczyk & Bertoncini, 1988; Gildea & Jurafsky, 1996; Plunkett, 1997; Bloom, 2010).

Assumption 2.2 Phonological processes only involve two levels of representation: the underlying, phonemic level and the surface, phonetic level (cf. Koskenniemi, 1983; Kaplan & Kay, 1994; Coleman, 1998).

Assumption 2.3 Phoneme-sized units are employed at all stages of phonological encoding and decoding (cf. Ohala, 1997).

Assumption 2.4 Infants are innately able to segment the stream of speech into phoneme-sized units (cf. Peperkamp et al., 2006; Le Calvez, 2007).

Assumption 2.5 Infants are good statistical learners and are able to monitor, for example, phone frequencies or transition probabilities between phones (cf. Saffran et al., 1996; Aslin et al., 1998; Saffran, 2002; Kuhl, 2004).

Assumption 2.6 Allophony and coarticulation processes can be reduced to strictly local, sound-altering processes that yield no segmental insertion or deletion (cf. Le Calvez, 2007; Goldsmith & Xanthos, 2009).

Assumption 2.7 Infants are able to undo (some) phonological variation without having yet acquired an adult-like lexicon (cf. Peperkamp & Dupoux, 2002; Peperkamp et al., 2006; Martin et al., 2009).

Assumption 2.8 Infants are able to acquire the phonemic inventory of their native language without knowledge of any other phonological construct such as features, syllables, or foots (cf. Le Calvez, 2007; Goldsmith & Xanthos, 2009; Boersma, 2010).

CHAPTER 3
SOURCES OF DATA

The aim of this chapter is to present the corpus of Japanese used throughout the present study, and to report all preprocessing steps applied to this master dataset. We also give an overview of the principles involved in deriving the allophonic rules whose modeling and acquisition are discussed in subsequent chapters. The focus is on presenting the similarities and discrepancies between our data-driven allophonic rules and traditional, theoretical descriptions of allophony in Japanese. It is worth highlighting immediately that, contrary to other chapters, the methodology discussed and used in this chapter is to a certain extent specific to the particular dataset we used: replacing this corpus with another one should yield no loss in the generality of the models to be further described except, obviously, with regard to conclusions drawn from quantitative or qualitative evaluations of the models' performance.

This chapter is divided into three main sections. Section 3.1 contains a general description of the corpus we used. In Section 3.2, we discuss how we derived allophones and allophonic rules from this corpus, as well as the mechanisms we used to control the limits of the phonetic similarity between the allophones of a given phoneme. Finally, Section 3.3 contains an examination of how our data eventually relate to theoretical descriptions of the phonology of Japanese.

3.1 The Corpus of Spontaneous Japanese

The corpus we used throughout this study to develop and validate our models is the *Corpus of Spontaneous Japanese* (henceforth CSJ; Maekawa et al., 2000; Maekawa, 2003), a large-scale annotated corpus of Japanese speech. The whole CSJ contains about 650 hours of spontaneous speech (corresponding to about 7 million word tokens) recorded using head-worn, close-talking microphones and digital audio tapes, and down-sampled to 16 kHz, 16 bit accuracy (CSJ Website, 2012). There is a true subset of the CSJ, referred to as the *Core*, which contains 45 hours of speech (about half a million words); it is the part of the corpus to which the cost of annotation was concentrated and, as a matter of fact, the only part where segment labels (i.e. phonetic and phonemic annotations) are provided. Therefore, from this point on, all mentions of the CSJ refer to the *Core* only.

On using the CSJ Although, for the sake (no pun intended) of brevity, we only report experiments using Japanese data in the present study, our goal is to develop a language-independent (dare we say, universal) computational model of the early acquisition of allophony and phonemehood. In other words, the models to be presented in subsequent chapters were not tuned to this specific corpus of Japanese, nor to any property of the Japanese language. Despite the fact that it has been argued, including by us, that computational models of early language acquisition should be evaluated using data from typologically different languages in order to assess their

sensitivity to linguistic diversity (Gambell & Yang, 2004; Boruta et al., 2011), we leave this as a recommendation for future research.

A more arguable choice was to use adult-directed speech as input data for models of early language acquisition. Notwithstanding the ongoing debate about how infant-directed speech might differ from adult-directed speech (Kuhl, 2004; Cristia, 2011), using infants' linguistic input to model infants' linguistic behavior would have been a natural choice. However, in the case at hand, our choice was constrained by the limited availability of transcribed speech databases. Indeed, to our knowledge, no database of infant-directed speech with aligned phonemic transcriptions was available in 2009, when the work reported in the present study was initiated. In any case, we follow Daland & Pierrehumbert's (2011) discussion, presented hereafter, concerning the relevance of using adult-directed speech as input data to models of early language acquisition. Would infant-directed speech be no different from adult-directed speech, then all conclusions drawn from the models' performance on the CSJ could be drawn for the acquisition of Japanese without loss of generality. On the contrary, would infant-directed speech indeed be different from adult-directed speech, and because studies supporting this alternative have argued that infant-directed speech is hyperarticulated in order to facilitate language acquisition (Kuhl, 2004; and references therein), then our models would have been evaluated in a worst case scenario.

A word about reproducibility Unfortunately, the CSJ is not freely available. Nonetheless, in order to guarantee the reproducibility of the work reported in this study, all programs and derivative data were made available to the community on a public repository at http://github.com/lucboruta.

Our concern for reproducibility also dictated the level of detail in this chapter, as well as our decision to report preprocessing operations from the beginning of this study, rather than in an appendix. Indeed, as will be further mentioned, ambiguous linguistic compromises or unspecified implementation choices have proven to hinder the repeatability of and comparability with previously reported experiments (viz. Dautriche, 2009; Martin et al., 2009).

3.1.1 Data preprocessing

Speech data in the CSJ were transcribed and annotated on various linguistic levels including, but not limited to, phonetics, phonology, morphology, syntax, and prosody. As the present study focuses on lower linguistic levels, the rich XML-formatted transcripts of the CSJ were preprocessed to extract information about phonemes, words, and (loosely defined) utterances. In order to train sound and accurate acoustic models, the focus of the preprocessing operations was on removing non-speech passages, as well as keeping track of temporal annotations so that the modified transcripts could still be aligned with the original audio recordings. Moreover, it is worth mentioning that although Dautriche (2009), too, used the CSJ, our data were extracted from newer, updated transcripts of the corpus.

Phonemes As previously stated, we want to address the issue of the limits of phonetic similarity, that is to say we want to control how varied and detailed our phonetic input will be. To do so, we follow a method initiated by Peperkamp et al. (2006) that we further systematized (Boruta, 2009, 2011a,b; Boruta et al., 2011): we created phonetic transcriptions by applying allophonic rules to a phonemically transcribed corpus. The particular technique we used in this study is presented in Section 3.2; here, our goal is to remove all phonetic annotations in the CSJ in order to obtain our initial phonemically transcribed corpus.

In the XML markup of the corpus, segmental data were annotated using two distinct elements: `Phone` and `Phoneme`. However, despite the names, a `Phoneme` element is made of one or more embedded `Phone` elements, and `Phone` elements may actually describe subsegmental events such as creakiness at the end of a vowel, or voicing that continues after the end of a vowel's formants (CSJ DVD, 2004). Therefore, we extracted the information about a phone's underlying phonemic

category from the PhonemeEntity attribute of the Phoneme element. As Phoneme elements do not include temporal annotations, we derived the timestamps of a phoneme from the timestamps of the contained Phone elements: the extracted starting timestamp of a phoneme is the starting timestamp of its first Phone and, mutatis mutandis, the ending timestamp of a phoneme is the ending timestamp of its last Phone.

Words The word segmentation and part-of-speech analyses of the CSJ were conducted for two different kinds of words: short-unit words (annotated as SUW) and long-unit words (annotated as LUW), which are defined as follows (CSJ Website, 2012):

> ❝Most of the SUW are mono-morphemic words or words made up of two consecutive morphemes, and approximate dictionary items of ordinary Japanese dictionaries. LUW, on the other hand, is for compounds.❞

Therefore, what will from this point on be referred to as words were extracted from the SUW tags. Moreover, the content of a given word was made up as the sequence of all embedded Phoneme elements, ignoring intermediate XML elements such as TransSUW or Mora, if any.

Utterances In the CSJ annotations, utterances are defined as inter-pausal units (tagged as IPU). According to the documentation, an utterance is a speech unit bounded by pauses of more than 200 ms (CSJ DVD, 2004). Nonetheless, extracting utterances and utterances' content is not a straightforward process as Noise tags were used as that level, too, to annotate unintelligible passages as well as anonymized data. Thence, an IPU element can be broken down into a sequence of one or more chunks; each chunk being either noise or workable words. Further annotation of such noise chunks in terms of words and phonemes is often incomplete, if not inexistent. It is also worth noting that, in the recordings, the audio passages matching some Noise elements were covered with white noise, so that looking past Noise tags (as did Dautriche, 2009) would surely compromise the training of acoustics-based models. For these reasons, all Noise chunks were discarded during preprocessing. For instance, an hypothetical IPU of the form `<IPU><SUW>...</SUW><Noise>...</Noise><SUW>...</SUW><SUW>...</SUW></IPU>` would be extracted as two distinct utterances: one containing only the first word, the other containing the two words appearing after the Noise chunk (unfortunately, the XML annotations are too rich and verbose so that an actual example could be presented on a single page).

In order to detect inconsistent annotations and to ensure the quality and the coherence of the extracted corpus, we performed various sanity checks, discarding utterance chunks that did not meet the following criteria:

— utterance chunks must contain at least one word;
— words must contain at least one phoneme;
— within a word, the ending timestamp of a phoneme must match the starting timestamp of the following phoneme;
— phonemes' timestamps must denote strictly positive durations;
— utterance chunks must be at least 100 ms-long, a threshold suggested by Dupoux & Schatz (priv. comm.).

Further modifications were applied, on the basis of the labels of some phoneme-level units; they are reported in the following section, together with a summary of all deletions in Table 3.1.

3.1.2 Data-driven phonemics

Completing the preprocessing operations described in the previous section yields a corpus reduced to three-level data: utterances that are made up of words that are made up of phonemes. Additional preprocessing is however necessary in order to obtain true phonemic transcriptions. The reason for this is twofold. On one hand, the categories gathered from the PhonemeEntity attributes do not only denote phonemic contrasts, but also allophonic contrasts. On the other hand, some phonemes of the inventory used in the CSJ are heavily under-represented and, in

Table 3.1 — Absolute and relative durations of retained and discarded data from the CSJ (159,965 seconds of recorded speech in total). Relative durations do not sum to 100% because of rounding.

	Absolute duration	Relative duration
Retained data	117,216	73%
Noisy or empty chunks	40,960	26%
Inconsistent or rare segments	1769	1%
Chunks less than 100 ms-long	16	$< 1\%$
Inconsistent timestamps	4	$< 1\%$

the perspective of conducting statistical or distributional learning, one can hardly have high expectations for a computational model if the occurrence of some target classes is anecdotal in its input. In this section, we discuss how we adapted the original transcription scheme of the CSJ to a novel, phonologically- and statistically-sound scheme.

Surprisingly, the theoretical descriptions of the phonology of Japanese we consulted proved to be divergent, if not contradictory. According to Ito & Mester (1995; incidentally, an explicit description of the phonemic inventory of Japanese is absent from their study), a proper description of the phonology of Japanese would require

> ❝the large-scale stratification of the Japanese lexicon into different classes [as] several phonological constraints are stratum-specific and hold only for a particular morpheme class.❞

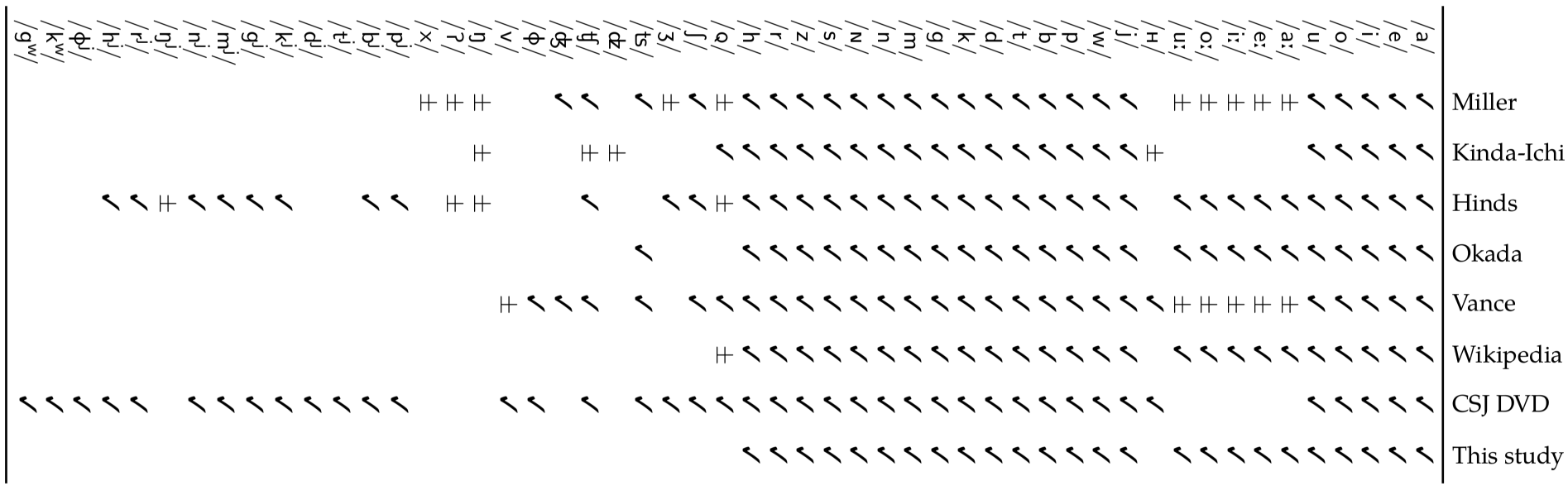
Here, we ignore such refinements and present a minimally sufficient description of the phonemic inventory of Japanese, designed as a trade-off between various theoretical phonological descriptions (Miller, 1967; Kinda-Ichi & Maës, 1978; Hinds, 1990; Ito & Mester, 1995; Okada, 1999; Vance, 2008; Wikipedia, 2012) and the annotation scheme used to tag the CSJ (CSJ DVD, 2004). The phonemic inventories used in each of the aforementioned references are presented and compared in Table 3.2, along with the final inventory used in this study. Let us mention that—for the purpose of the present analysis—we use the terms *syllable* and *mora* interchangeably. It is also worth highlighting that non-local aspects of the phonology of Japanese were ignored, most notably *rendaku* (Otsu, 1980; Vance, 1980) and pitch accent (e.g. Hinds, 1990; Okada, 1999).

Garbage in, garbage out The original inventory of the CSJ contains a handful of unworkable phoneme-level units whose utterances were merely discarded from the corpus. First, although the segment `[VN]` occurs 1712 times in `PhonemeEntity` attributes, this tag was used to annotate filler words and hesitations (CSJ DVD, 2004). For this reason, it can not reasonably be mapped to any phoneme, and the utterances in which it occurs were discarded. Following the same argument, any utterance containing the tags `[?]` or `[FV]`, respectively used throughout the corpus to annotate unidentifiable consonants and vowels (1214 occurrences in total), were deleted. Ultimately, utterances containing heavily under-represented phoneme-level units such as `[kw]`, `[Fy]`, and `[v]` (1, 1, and 2 occurrences, respectively) were also discarded with no further examination of their actual phonemic status.

This final data-discarding step accounts for a very small fraction of all discarded data. Indeed, inclusion and deletion statistics presented in Table 3.1 indicate that discarded data consists, for the most part, in noisy or empty utterance chunks. Having noted that Dautriche (2009) did not account for noisy passages, we are puzzled by the very good results reported in this study.

Vowels The annotation of vowels in the CSJ follows Vance's (2008) description: there are five vowel qualities and one chroneme (i.e. a phoneme-level unit conveying only length information) in Japanese. Each vowel quality was annotated with one unambiguous tag: `[a]`, `[e]`, `[i]`, `[o]`, and `[u]`. Thus, in isolation, the CSJ tag `[a]` was straightforwardly mapped to the phoneme /a/ and, similarly, `[e]` to /e/, `[i]` to /i/, `[o]` to /o/, and `[u]` to /u/ (we consider the fact that realizations of this last phoneme are typically unrounded [ɯ] irrelevant for the purpose of the

Table 3.2 — Comparison of various descriptions of the phonemic inventory of Japanese (Miller, 1967; Kinda-Ichi & Maës, 1978; Hinds, 1990; Okada, 1999; CSJ DVD, 2004; Vance, 2008; Wikipedia, 2012; the present study). Although aforementioned references use disparate notational systems, or different symbols for the same phoneme, all inventories are presented here using the symbols of the international phonetic alphabet, notwithstanding the moraic phonemes /ǫ/ and /ʜ/. A checkmark indicates that the phoneme is included in the author's phonemic inventory. A plus-minus sign indicates that the author discusses the phonemic status of the segment, but does not include it in the phonemic inventory.



description of a phonemic inventory and, therefore, use the usual symbol for the cardinal vowel).

The moraic chroneme tag [H] was used to annotate the second half of a long vowel and, hence, serves as an abstract placeholder for the length mark /ː/ in any of /aː, eː, iː, oː, uː/. In the perspective of training acoustics-based models, the problem with such a segment is that its realizations would encompass all possible vowel qualities. Therefore, in order to set all long vowels apart, this tag and the preceding one were mapped to a single long vowel whose quality is described by the preceding tag. For instance, the sequences [aH], [eH], [iH], [oH], and [uH] were mapped to /aː/, /eː/, /iː/, /oː/, and /uː/, respectively. Because, in terms of segments, this mapping amounts to deleting the moraic chroneme, each long vowel's timestamps were corrected, extending the ending timestamp of the vowel to the ending timestamp of the chroneme.

Plain consonants Each consonant's default realization was annotated with an unambiguous tag. Hence, we straightforwardly mapped these tags to their respective phonemes: [p] to /p/, [t] to /t/, [k] to /k/, [b] to /b/, [d] to /d/, [g] to /g/, [m] to /m/, [n] to /n/, [r] to /r/, [h] to /h/, [s] to /s/, and [z] to /z/. Likewise, the approximants received no specific treatment: the tags [w] and [y] were mapped to /w/ and /j/, respectively. Finally, following all aforementioned grammars of Japanese, the moraic nasal tag [N] was mapped as-is onto the phoneme /ɴ/.

Glottal fricatives The tagset of the CSJ uses the symbol [F] to represent a voiceless bilabial fricative. This segment is however not a phoneme in Japanese, but the allophonic realization [ɸ] of the glottal fricative /h/ before /u/ (Kinda-Ichi & Maës, 1978; Okada, 1999; CSJ DVD, 2004). Therefore, both [h] and [F] tags were mapped to the phoneme /h/.

Yōon and palatalizations *Yōon*, literally "contracted sound" or "diphthong," is a feature of the phonology of Japanese in which a mora is formed with an added yod after the initial consonant. Although actual realizations of such sequences result in the palatalization of the initial consonant, *yōon* should not be confused with allophonic palatalizations triggered by a following front vowel. Indeed, as exemplified by the minimal pair ⟨ko⟩ ∼ ⟨kyo⟩, "child" ∼ "hugeness" (we purposedly use angle brackets to temporarily sidestep the issue of the phonemic representation), *yōon* and plain consonants form distinct and contrastive groups for Japanese speakers (Nakamura–Delloye, priv. comm.). In the CSJ, each consonant may receive one of three atomic tags: plain consonant, e.g. [k] and [n], palatalized consonant, [kj] and [nj], or *yōon*, [ky] and [ny]. Although, as it is a phonemic contrast, distinguishing *yōon* from plain consonants is legitimate, tagging allophonic palatalizations as phonemes is not. The annotation of consonantal phonemes was thus simplified to account for phonemic contrasts only: both [k] and [kj] were mapped to the phoneme /k/ and, similarly, [g] and [gj] to /g/, [n] and [nj] to /n/, and [h] and [hj] to /h/; no allophonic palatalization of /p/, /b/, /t/, /d/, /m/, or /r/ was tagged as such in the transcripts.

As for the *yōon*, although they can not be mapped to plain phonemes, it should be noted that they occur quite infrequently compared to their plain counterparts. Indeed, tallying occurrence frequencies over all the transcripts of the CSJ, only 283 segments were tagged as [py] vs. 4453 as [p], 603 as [ty] vs. 96903 as [t], 3186 as [ky] vs. 99170 as [k] or [kj], 234 as [by] vs. 13061 as [b], 10 as [dy] vs. 47447 as [d], 598 as [gy] vs. 31416 as [g] or [gj], 1187 as [ry] vs. 57640 as [r], 81 as [my] vs. 54932 as [m], 400 as [ny] vs. 91666 as [n] or [nj], and 1282 as [hy] vs. 21285 as [h], [hj], or [F]. Hence, the ratios of *yōon* to plain consonant range from 1:16 to 1:4748. However, in the interest of conducting statistical learning, it is desirable to have a reasonably high number of examples for each phoneme and, thus, we need to attenuate the potential data-sparseness of our dataset. Therefore, and although such segments were considered as atomic in the annotations of the CSJ, all *yōon* were mapped to a sequence of two phonemes consisting of the plain version of the consonant and a yod; i.e. the atomic tag [py] was mapped the bigram /pj/ and, similarly, [ty] to /tj/, [ky] to /kj/, [by] to /bj/, [dy] to /dj/, [gy] to /gj/, [ry] to /rj/, [my] to /mj/, [ny] to /nj/, and [hy] to /hj/. As a consequence, a new timestamp, marking the boundary between the consonant and the yod, had to be created for each bigram. This additional timestamp was

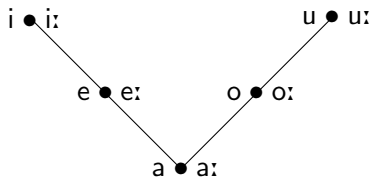


Figure 3.1 — Vowels of the phonemic inventory of Japanese, as used in this study. Symbols at left and right of bullets represent short and long vowels, respectively.

Table 3.3 — Consonants and glides of the phonemic inventory of Japanese, as used in this study. Where symbols appear in pairs, the one to the right represents a voiced consonant.

	Bilabial	Alveolar	Palatal	Velar	Uvular	Glottal
Nasal	m	n			ɴ	
Plosive	p b	t d		k g		
Fricative		s z				h
Flap		r				
Approximant			j	w		

computed pro rata temporis of the average observed durations of the standalone occurrences of the plain consonant and the yod.

Yod-coalescence Alveolar fricative consonants received a separate treatment from other consonants. Based on preliminary observations by Dupoux & Schatz (priv. comm.) and our own examination of approximately 50 randomly-drawn phones, we observed that, for each of these consonants, their default realization, allophonic palatalization, and *yōon* resulted in a common, perceptually atomic phone. Therefore `[s]`, `[sj]`, and `[sy]` were mapped to the phoneme /s/. Following the same argument, `[z]`, `[zj]`, and `[zy]` were mapped to the phoneme /z/. As far as *yōon* is concerned, this observation is not without reminding us of yod-coalescence, a process occurring in many varieties of English (Wells, 1999) by which the clusters [dj], [tj], [sj], and [zj] are changed into [ʤ], [ʧ], [ʃ], and [ʒ], respectively. Moreover, `[c]`, `[cj]`, and `[cy]` were all mapped to the phoneme /t/ on the grounds that [c] does not occur in the phonology of standard Japanese, except as the allophonic realization of /t/ before /u/ (Ito & Mester, 1995; p. 825).

Moraic obstruents Finally, the moraic obstruent tag `[Q]` was used throughout the CSJ to annotate the second half of a geminate obstruent and, hence, serves as an abstract placeholder for any of /p, t, k, b, d, g, s, z, h/, depending on the context. In order to set all obstruents apart, and following our discussion on the moraic chroneme tag `[H]`, this tag was mapped to the phoneme to which the preceding segment had been mapped. When the preceding segment was not an obstruent, the moraic segment was used to annotate an emphatic pronunciation, as described by Hinds (1990; p. 399). In such contexts, this segment was simply deleted, thereby extending the ending timestamp of the preceding segment to the ending timestamp of the moraic obstruent.

3.1.3 The phonemic inventory of Japanese

After applying all aforementioned preprocessing steps, the phonemic corpus we derived for our experiments is very different from the original CSJ. Because the phonemic inventory of this novel corpus will be at the heart of most if not all subsequent discussions, we give a brief overview of the comprised phonemes in this section.

Let the set $\mathfrak{P}$ denote the resulting phonemic inventory of Japanese, we have:

$$\mathfrak{P} \equiv \{\text{a}, \text{e}, \text{i}, \text{o}, \text{u}, \text{aː}, \text{eː}, \text{iː}, \text{oː}, \text{uː}, \text{j}, \text{w}, \text{m}, \text{n}, \text{ɴ}, \text{p}, \text{b}, \text{t}, \text{d}, \text{k}, \text{g}, \text{s}, \text{z}, \text{h}, \text{r}\} \tag{3.1}$$

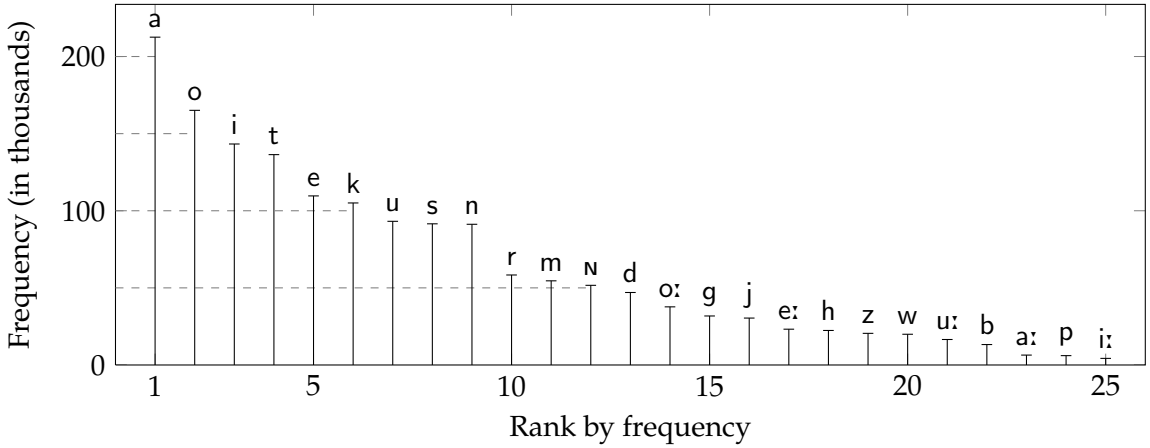


Figure 3.2 — Distribution of phoneme frequencies observed in the CSJ, as a function of phoneme ranks in the frequency table.

This inventory is also presented in Figure 3.1 for vowels and in Table 3.3 for consonants, using the usual representations established in the charts of the international phonetic alphabet (International Phonetic Association, 1999).

In adapting the annotations of the CSJ to our needs, we intentionally made a trade-off between linguistic and computational considerations. Whereas we examined all aforementioned theoretical descriptions of the phonology of Japanese with equal interest, it is interesting to observe that our inventory is (but for one phoneme in each case) identical to those proposed by, on one hand, an expert linguist and native speaker (Okada, 1999) and, on the other hand, a collaboratively edited Internet encyclopedia (Wikipedia, 2012).

Power phonemes One of our recurrent concerns during the preprocessing phase was to prevent potential data-sparseness issues in our corpus. To do so, we mapped the annotations of the CSJ to a minimal number of phonemes, thereby increasing the frequency of the latter. Nonetheless, it has often been observed that the frequency distributions of many linguistic units follow a power law (i.e. a few items have very high frequencies and most items have very low frequencies; Zipf, 1935; Ferrer-i-Cancho & Solé, 2001, 2003; Ha et al., 2002; Tambovtsev & Martindale, 2007), and the frequency distribution of phonemes in our corpus is no exception.

Indeed, as presented in Figure 3.2, phonemes' absolute frequencies as tallied over the whole CSJ range from 4280 occurrences for /iː/ to 212,554 occurrences for /a/. In other words, the most frequent phoneme in our corpus is 50 times more frequent than the least frequent phoneme. As will be discussed in Chapter 4, it would be unrealistic to have equal expectations for the most and the least frequent phonemes in the perspective of conducting statistical learning on such a dataset; yet, while power laws appear to be an intrinsic property of any natural language, they are not a curse for computational approaches to language and we will discuss how introducing a weighting scheme in our experiments may account for this fact.

3.2 Good allophones like yo momma used to cook

In this section, we describe how we derived phonetic transcriptions with varying degrees of allophonic variability using both the original audio recordings and our novel phonemic transcription of the CSJ. We first give a brief overview of similar techniques used in previously reported studies, and we then proceed to detail the particular method we used in this study, and discuss how it allows us to control the limits of phonetic similarity in our data.

As previously stated, our goal is to create a phonetically transcribed corpus from a phonemically transcribed one, controlling the amount of allophonic variation we introduce. The reason for

this is twofold. First, the number of allophonic rules infants must learn is unknown (if assessable at all; Boruta, 2011b) and, thus, exploring a considerable range of inputs of increasing phonetic variability can yield but a better evaluation of our models' performance. Second, as we build upon Peperkamp et al.'s early studies, we need input data comparable to the corpora they used, i.e. data from which statistics about each phone's distribution can be gathered. Therefore, we can not use raw speech as it would make the collection of such distributional statistics impossible: if no two phones are identical, monitoring the occurrences of a given phone in different contexts becomes impractical. In other words, Peperkamp et al.'s (2006) distributional learner requires that the input be quantized, meaning that the virtually infinite variability in human speech and phoneme realizations has to be somehow constrained to a finite (and relatively small) number of values, that we will refer to as the *allophonic inventory* denoted by P:

Assumption 3.1 Infants' linguistic input is (perceptually) quantized, i.e. there is a finite number of phones to keep track of.

Such a preprocessed representation of the input is comparable (yet unrelated) to McCallum et al.'s (2000) canopy clustering whereby, in order to facilitate the application of clustering techniques on datasets of considerable size, an initial, relatively fast or approximate preclustering is performed to partition the data into a restricted number of subsets—referred to as canopies—allowing for the subsequent use of resource-intensive clustering techniques on the makeshift canopies. For the purpose of satisfying this assumption, Peperkamp et al. (2006; and subsequent studies) applied allophonic rules to phonemically transcribed corpora. Doing so, they also made the simplifying assumption that free variation is beyond the scope of their research:

Assumption 3.2 There can be no more than one phone per phoneme per context.

Building upon this framework, most subsequent studies (Dautriche, 2009; Martin et al., 2009; Boruta, 2009, 2011a,b; Boruta et al., 2011) made explicit the additional simplifying assumption that phonemes do not have common realizations, predominantly in order to avoid frequency or probability mass derived from different phonemes merging onto common, indistinguishable phones (Boruta et al., 2011; p. 4):

Assumption 3.3 Phonemes do not have common realizations, i.e. the phonemic inventory is a partition of the allophonic inventory.

For the sake of comparability, we reiterate these assumptions in the present study. The remainder of this section contains a brief review of the different techniques used in aforementioned experiments to transform phonetic transcriptions into phonemic transcriptions.

Allophonic complexity The statistic we use troughout the present study to quantify how far the allophonic inventory is from the optimal phonemic inventory is the *allophonic complexity* of the corpus (Boruta, 2009, 2011a,b; Martin et al., 2009; Boruta et al., 2011). It is defined as the average number of allophones per phoneme, i.e. the ratio of the number of phones to the number of phonemes in the language. Let $\mathfrak{n} \equiv |\mathfrak{P}|$ be the number of phonemes in the language at hand (viz. $\mathfrak{n} = 25$ in Japanese) and n be the number of phones in a given allophonic inventory, the allophonic complexity of the corresponding corpus is simply given by $n/\mathfrak{n}$.

In Table 3.4, we present a summary of the various allophonic complexities, languages, corpora, and (looking ahead) indicators of allophony with which experiments building upon Peperkamp et al.'s (2006) study were conducted.

3.2.1 Old recipes for allophonic rules

Different techniques have been used to convert a phonetically transcribed corpus into a phonemically transcribed one. This section offers a brief overview of their respective advantages and limitations.

Table 3.4 — Summary of the parameters used in computational studies building upon Peperkamp et al.'s (2006) study. In the second column, ID and AD stand for infant-directed speech and adult-directed speech, respectively. For each study, used indicator classes are denoted by their initial: $\mathbb{A}$ for acoustic, $\mathbb{D}$ for distributional, $\mathbb{L}$ for lexical, $\mathbb{P}$ for phonetic (articulatory), and $\mathbb{T}$ for temporal indicators. The *and/or* conjunction is used for studies in which the problem of combining different indicators is addressed. In the last column, checkmarks denote experiments whose evaluation was a true evaluation, rather than a prognosis.

Study	Data	Indicator classes	Allophonic comp.	Eval.
Peperkamp et al. (2006)	Artificial	$\mathbb{D}$	$n/n = 47/46$	
	ID French	$\mathbb{D}$ and $\mathbb{P}$	$n/n = 44/35$	✓
Le Calvez (2007)	Artificial	$\mathbb{D}$	$61/60 \leq n/n \leq 95/60$	✓
	ID English	$\mathbb{D}$	$n/n = 51/50$	✓
	ID English	$\mathbb{D}$ and $\mathbb{P}$	$n/n = 136/50$	✓
	ID Japanese	$\mathbb{D}$	$n/n \in \{50/49, 46/42\}$	✓
	ID Japanese	$\mathbb{D}$ and $\mathbb{P}$	$n/n = 58/42$	✓
	ID French	$\mathbb{D}$	$n/n \in \{36/35, 43/35\}$	✓
	ID French	$\mathbb{D}$ and $\mathbb{P}$	$n/n = 45/35$	✓
Le Calvez et al. (2007)	Artificial	$\mathbb{D}$	$61/60 \leq n/n \leq 95/60$	
	ID French	$\mathbb{D}$ and $\mathbb{P}$	$n/n = 45/35$	✓
	ID Japanese	$\mathbb{D}$ and $\mathbb{P}$	$n/n = 53/49$	✓
Martin et al. (2009)	AD Japanese	$\mathbb{D}$	$79/42 \leq n/n \leq 737/42$	
	AD Japanese	$\mathbb{D}$ and $\mathbb{L}$	$79/42 \leq n/n < 1800/42$	
	AD Dutch	$\mathbb{D}$ and $\mathbb{L}$	$93/50 \leq n/n < 2200/50$	
Boruta (2009)	ID English	$\mathbb{D}$ and $\mathbb{L}$	$99/50 \leq n/n \leq 1136/50$	
	ID French	$\mathbb{D}$ and $\mathbb{L}$	$69/35 \leq n/n < 781/35$	
	ID Japanese	$\mathbb{D}$ and $\mathbb{L}$	$94/49 \leq n/n \leq 558/49$	
Dautriche (2009)	AD Japanese	$\mathbb{A}$, $\mathbb{D}$, and $\mathbb{P}$	$100/49 \leq n/n \leq 1000/49$	
Boruta (2011b)	ID English	$\mathbb{D}$ and/or $\mathbb{L}$	$99/50 \leq n/n \leq 1136/50$	✓
This study	AD Japanese	$\mathbb{A}$, $\mathbb{D}$, $\mathbb{L}$, and/or $\mathbb{T}$	$48/25 \leq n/n \leq 990/25$	✓

Hand-crafted rules In their first studies, Peperkamp et al. created phonetic corpora through the automatic application of hand-written allophonic rules (Peperkamp et al., 2006; Le Calvez, 2007; Le Calvez et al., 2007). This method has two obvious advantages as the rules they used are not only linguistically plausible, but also attested in the language at hand. However, and notwithstanding the fact that this procedure is not fully automatic, the reduced allophonic grammars they defined introduced a limited number of contextual variants, as reported in Table 3.4.

Random-partition rules Martin et al. (2009) circumvented this limitation using a language-independent algorithm whereby an allophone is simply a numbered version of the underlying phoneme, e.g. using [p$_1$], [p$_2$], [p$_3$] as the contextual realizations of /p/. In practice, the allophones of a given phoneme are generated by partitioning the set of this phoneme's possible contexts (i.e. the phonemic inventory of the language at hand) in as many subsets as necessary to reach a desired allophonic complexity. However, because of Martin et al.'s choice to generate allophones using the whole phonemic inventory rather than each phoneme's attested contexts, this procedure may yield inapplicable rules, e.g. describing the realization of /w/ between /ŋ/ and /k/ in English, thus artificially limiting the maximum attainable allophonic complexity.

Matrix-splitting rules In a previous study (Boruta, 2009, 2011a; Boruta et al., 2011), we presented an algorithm that combines the linguistic plausibility of Peperkamp et al.'s rules and

Martin et al.'s ability to emulate rich transcriptions. Relying on a representation of the language's phonemes in terms of distinctive features à la Chomsky & Halle (1968), this algorithm iteratively splits phonological matrices, updating them with distinctly valued attributes one at a time so that, in the end, a phonological grammar of assimilation rules is produced (Boruta, 2011a). Finally, a phonetic transcription is created by applying this artificial grammar to a phonemic transcription. However, it is worth noting that this algorithm gives nothing but an approximation of what speech-based allophones might sound like. It was designed to mimic the algorithm described in the following section when only transcriptions (and no speech recordings) are available.

3.2.2 Allophonic rules redux

Using the CSJ does not only allow us to explore a linguistic database of considerable size. Because the content of each transcript in the corpus is aligned with the original audio recordings, using the CSJ allows us to ground our experiments in the acoustic form of language, thus bridging to some extent the gap between phonology-driven and phonetics-driven approaches to the acquisition of phonological knowledge.

The main idea behind the technique we present in this section is to create, for each phoneme, an acoustic model of all attested realizations, controlling how much information about contextual variability is preserved. It is worth mentioning immediately that the methodology we adopted is not ours: we follow the protocol described by Young et al. (2006; section 3), as implemented by Dautriche (2009; pp. 11–17) and Dupoux & Schatz (priv. comm.) using the *Hidden Markov Model Toolkit* (henceforth HTK; Young et al., 2006), albeit while adapting all linguistic aspects to the case at hand. Although we adopted this protocol as-is, we will hereby describe the core concepts of the procedure, as well as the implementation choices specific to the present study.

Coding the data Discussing the details of automatic speech processing is beyond the stated scope of this research; suffice it to say that the following protocol describes a now-standard analysis of speech recordings (Makhoul & Schwartz, 1995; Hermansky, 1999; Young et al., 2006; Jurafksy & Martin, 2009). The initial step consists in transforming the raw speech waveform of the CSJ recordings into sequences of numerical vectors, a.k.a. acoustic features. A schematic view of this process is presented in Figure 3.3: acoustic features are computed from successive and overlapping slices of the speech waveform. Concretely, acoustic features for our experiments

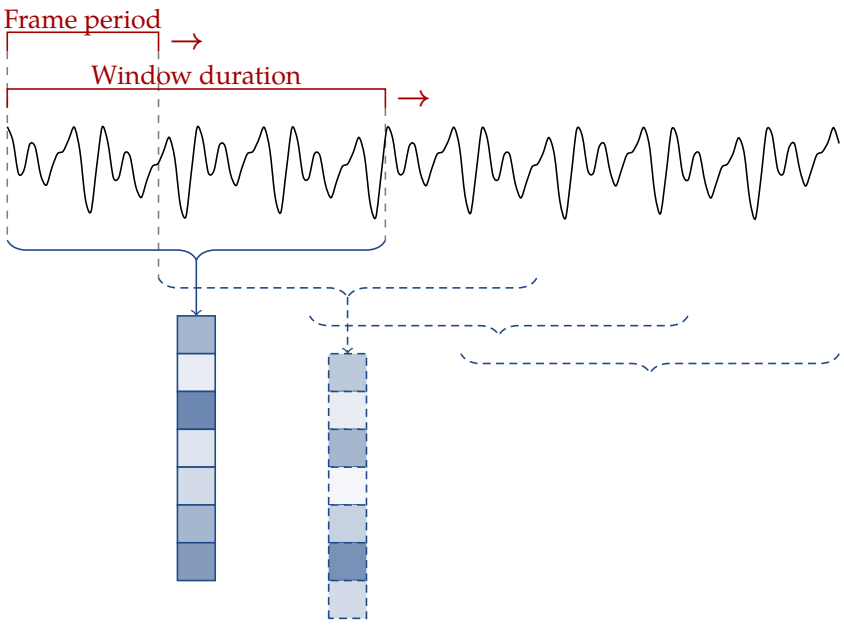


Figure 3.3 — Schematic representation of acoustic feature extraction by short-time analysis.

Table 3.5 — Configuration parameters for the extraction of MFCC vectors. Checkmarks indicate default values in HTK 3.4.1 (Young et al., 2006), given here for completeness.

Parameter	= Value	
SOURCEFORMAT	= WAV	
TARGETFORMAT	= HTK	✓
TARGETKIND	= MFCC_0_D_A	
TARGETRATE	= 100000.0	
WINDOWSIZE	= 250000.0	
USEHAMMING	= T	✓
NUMCEPS	= 12	✓
CEPLIFTER	= 22	✓
PREEMCOEF	= 0.97	✓
NUMCHANS	= 26	
ENORMALISE	= F	
SAVEWITHCRC	= T	✓
SAVECOMPRESSED	= T	

were derived from a short-time analysis of the speech signal with a 25 ms Hamming window advanced in 10 ms steps (a.k.a. the frame period in HTK's jargon and Figure 3.3). For each slice, we extracted 12 Mel frequency cepstral coefficients (henceforth MFCC; e.g. Jurafksy & Martin, 2009; p. 295–302), along with an energy coefficient and the first- and second-order time derivative coefficients (the so called deltas and double-deltas) for each of the first 13 coefficients. For the sake of reproducibility, HTK's configuration parameters are reported in Table 3.5.

It is worth mentioning that 8 of the 201 recordings in the CSJ are actually dialogues, a specificity that seemingly was not taken into consideration by Dautriche (2009). Fortunately, in order to distinguish between both speakers, each of them was attributed one stereo channel in the audio recordings. Thus, based on the information available in the `Channel` attribute of the `IPU` elements, we were able to configure the extraction of the acoustic features for each speaker and utterance, matching the value of HTK's `STEREOMODE` parameter to the appropriate stereo channel.

Creating phonemic HMMs The next step in this protocol is to train, for each phoneme in the inventory, a model of all attested realizations in the corpus. In HTK, an acoustic model takes the form of a *Hidden Markov Model* (henceforth HMM; e.g. Young et al., 2006; p. 3). Suffice it to say that a HMM is a statistical model comprising a finite number of connected states, and that some if not all states (a.k.a. the emitting states) are associated to probabilities governing the distribution of observations (i.e. the acoustic features) at a particular time given the state of a hidden variable (i.e. the underlying phonemes) at that time. The general topology of the HMMs used in HTK is presented in Figure 3.4. Because this topology is unique and constant throughout the present

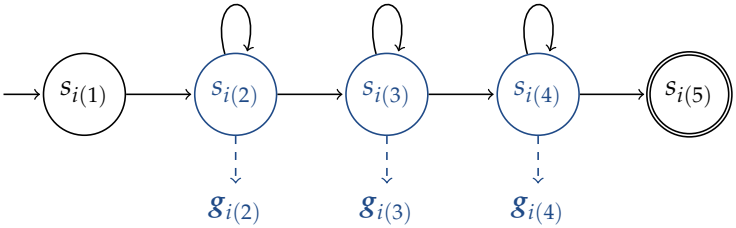


Figure 3.4 — HMM topology for the phoneme models, i.e. a 3-state left-to-right HMM with no skip. In HTK, the initial and accepting state are non-emitting; each emitting state $s_{i(x)}$ is associated to an output probability density function $g_{i(x)}$ (here in blue).

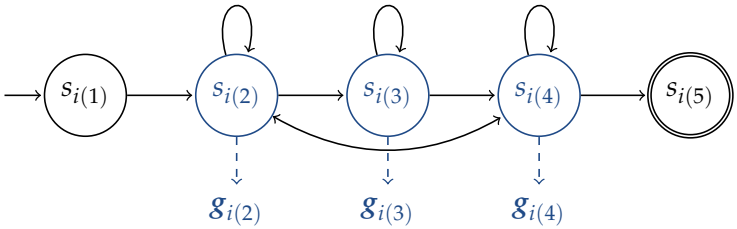


Figure 3.5 — HMM topology for the silence models trained for the boundaries <s> and </s>. The extra two-way transition between the second and the fourth states allows the model to absorb impulsive noises without committing to the next phone (Young et al., 2006; p. 32).

study, let $\mathbf{s}_i \equiv \langle s_{i(2)}, s_{i(3)}, s_{i(4)} \rangle$ denote the sequence of all emitting states in the HMM modeling the realizations of the phone $p_i \in P$, and $g_{i(x)}$ denote the output probability density function associated to the emitting state $s_{i(x)}$. Because silence or pauses may occur at utterance boundaries, two additional silence-based HMMs are also trained, viz. one for the starting boundary and one for the ending boundary, denoted <s> and </s>, respectively. The particular topology of these HMMs is illustrated in Figure 3.5.

In HTK, the output probability density function linked to each emitting state is represented by a Gaussian mixture model (Young et al., 2006; p. 6). Using Gaussian mixture models (henceforth GMM; e.g. Reynolds, 2009) is particularly appropriate to account for the inherent variability in human speech. Indeed, as emphasized by Goldberger & Aronowitz (2005):

> ❝the main attraction of the GMM arises from its ability to provide a smooth approximation to arbitrarily shaped densities of long term spectrum.❞

A GMM g is nothing but a parametric density function, defined as a weighted sum of $\gamma \in \mathbb{N}$ component Gaussian density functions. Concretely, the definition of the probability density function of a GMM is given by

$$g(\mathbf{x} \mid \Theta) \equiv \sum_{c=1}^{\gamma} \varpi_c \, g(\mathbf{x} \mid \theta_c) \tag{3.2}$$

where $\mathbf{x} \in \mathbb{R}^b$ is a b-dimensional random vector, $\Theta \equiv (\theta_1, \ldots, \theta_\gamma)$ contains the mixture's complete parameterization (to be defined hereafter), $\varpi_c \geq 0$ is the weighting factor for the c-th mixture component, and $g(\mathbf{x} \mid \theta_c)$ is the c-th component's probability density function. Each component's density function is a b-variate Gaussian function of the form

$$g(\mathbf{x} \mid \theta) \equiv \mathcal{N}_b(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \equiv \frac{\exp\left(-\left((\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu})\right)/2\right)}{\sqrt{(2\pi)^b \det(\boldsymbol{\Sigma})}} \tag{3.3}$$

where $\theta \equiv (\varpi, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the component parameters, among which $\boldsymbol{\mu}$ is the b-dimensional mean vector of the component and $\boldsymbol{\Sigma}$ is its $b \times b$ diagonal covariance matrix. The component weights satisfy the constraint that $\sum_{c=1}^{\gamma} \varpi_c = 1$ in order to ensure that the mixture is a true probability density function (Stylianou, 2008; Reynolds, 2009). In the case at hand, the dimensionality b matches the number of coefficients in each MFCC vector, i.e. $b = 39$ as the acoustic features were extracted using 12 cepstral coefficients, an additional energy coefficient, and the first- and second-order time derivatives for each of the first 13 coefficients. The default number of mixture components in HTK is $\gamma = 17$ (we were unable to find a reference justifying this ubiquitous value, reproduced here as-is from Young et al.'s and Dautriche's studies). Finally, it is worth noting that all parameter values collected in Θ are automagically estimated by HTK from its input data.

Creating allophonic HMMs The third step of this protocol consists in training allophonic HMMs, specificizing phonemic HMMs using information about each phoneme's attested context-dependent realizations. This is the step where we are able to control the size of the allophonic inventory. Using HTK, allophonic HMMs are initialized by simply cloning phonemic HMMs

and then retraining all models using so called triphone transcriptions (Young et al., 2006; p. 34–36), i.e. phonetic transcriptions where each phone is denoted by the trigram made up of the preceding phoneme, the target phoneme, and the following phoneme. For example, using HTK's minus-plus notation and visible spaces to denote segment boundaries, an utterance consisting of the word /sake/, would be transcribed as <s>-s+a␣s-a+k␣a-k+e␣k-e+</s> where s-a+k denotes the realization of the phoneme /a/ between /s/ and /k/—the treatment of the initial and final segments was simplified for the sake of the example. It is worth noting that such a transcription yields the highest attainable allophonic complexity for any corpus. Indeed, following Peperkamp et al. (2006; and subsequent studies), we assumed that there can be no more than one phone per phoneme per context (Assumption 3.2). In a triphone transcription, a phone is systematically represented by the trigram made up of the underlying phoneme and its attested context. Overall, the allophonic inventory in such a transcription comprises, for every phoneme in the language at hand, the set of all realizations in all attested contexts. Therefore, we can not further increase the allophonic complexity of our dataset while still satisfying the aforementioned assumption.

Having re-estimated the parameters of all allophonic HMMs, the next step consists in controlling the overall allophonic complexity by merging, for each phoneme, some of these utterly specified allophones—a procedure known as state-tying in HTK's jargon (Young et al., 2006; pp. 36–39, 144–146). For a given phoneme and a given emitting state position $x \in \{2, 3, 4\}$, the x-th states of all allophonic HMMs of this phoneme are first pooled into a single cluster, and then iteratively partitioned into an arborescent structure resembling a decision tree; this process is repeated for each phoneme and each of the three emitting states in the HMMs. The iterative partitioning relies on a set of a priori linguistic questions ranging from natural classes (e.g. consonants, vowels, stops) to specific contexts; the questions we designed for the purpose of the present study are reported in Table 3.6 where each set of contexts was used to create questions for both preceding and following contexts. For a given cluster, HTK applies each question in turn, tentatively splitting the cluster into two new branches comprising, on one hand, phones whose realization context belongs the question's application contexts (e.g. the phone's context is a voiced phoneme) and, on the other hand, phones whose realization context does not belong to the question (e.g. the phone's context is not a voiced phoneme). The question which maximizes the likelihood of the data—i.e. the probability of partitioning these given phones in this particular way given the training data—is then effectively used to split the initial cluster, and this process is repeated on every new branch until an a priori global threshold (to be discussed hereafter) is reached. In the end, for each leaf in this newly created decision tree, the HMM states within that leaf are said to be tied in the sense that, during a final re-estimation of all models' parameters, a single, common GMM is used for every set of tied states. The number of allophones that were computed for a given phoneme is defined as the sum, for each emitting state, of the number of distinct (i.e. not tied) leaves.

Once all state-tying has been completed, some allophones may share the exact same HMM and, in that case, the corresponding phones are indistinguishable (at least in our quantized representation). Following the work of Dautriche (2009) and Dupoux & Schatz (priv. comm.), our termination criterion for state-tying relies on the global number of such tied, indistinguishable phones: let $\tilde{n}$ denote the desired number of phones in the allophonic inventory, our solution consists in the inventory that maximizes the overall likelihood of the tied allophonic HMMs given the CSJ, and whose size n is as close as possible to $\tilde{n}$. Indeed, because of unspecified reasons, Dautriche's and Dupoux & Schatz's programs, though deterministic, do not always output an inventory comprising the exact desired number of phones; observed differences between the desired and the actual number of phones range from -18 to $+2$.

As previously stated, our goal is to explore a considerable range of inputs of increasing phonetic variability. In order to do so, we ran 20 distinct state-tying procedures controlling the a priori size of the allophonic inventory, ranging from 50 phones to 1000 phones in increments of 50. Whereas the lower bound and the increment were chosen arbitrarily, it should be noted that requesting allophonic complexities higher than $1000/25$ requires considerable computing time

Table 3.6 — Questions on contexts used in HTK's decision-tree state-tying procedure. All questions were input to to the system in the order specified in this table, duplicating each question in turn to account successively for phonemes' preceding and following contexts (except for the question on silence, because of the use of distinct boundary denotations).

	Question	Set of context phonemes
1	Voiced phoneme?	{a, e, i, o, u, aː, eː, iː, oː, uː, b, m, n, d, z, r, g, ɴ, j, w}
2	Consonant?	{p, b, m, t, d, n, s, z, r, k, g, j, w, ɴ, h}
3	Vowel?	{a, e, i, o, u, aː, eː, iː, oː, uː}
4	Plosive consonant?	{p, b, t, d, k, g}
5	Alveolar consonant?	{n, t, d, s, z, r}
6	Short vowel?	{a, e, i, o, u}
7	Long vowel?	{aː, eː, iː, oː, uː}
8	Voiceless consonant?	{p, t, k, s, h}
9	High vowel?	{i, u, iː, uː}
10	Back vowel?	{o, u, oː, uː}
11	Palatalization trigger?	{i, iː, j}
12	Nasal consonant?	{m, n, ɴ}
13	Bilabial consonant?	{m, p, b}
14	Velar consonant?	{k, g,w}
15	Fricative consonant?	{s, z, h}
16	Denti-alveolar consonant?	{n, t, d}
17	Glide?	{j, w}
18	Vowel quality [a]?	{a, aː}
19	Vowel quality [e]?	{e, eː}
20	Vowel quality [i]?	{i, iː}
21	Vowel quality [o]?	{o, oː}
22	Vowel quality [u]?	{u, uː}
23	Alveolar consonant?	{s, z}
24	Flap consonant?	{r}
25	Uvular consonant?	{ɴ}
26	Glottal consonant?	{h}
27	Silence?	{<s>} or {</s>}

and power, notwithstanding those necessary to conduct subsequent experiments. However, it is worth highlighting that the highest allophonic complexity that can reasonably be achieved with HTK's speech-based state-tying procedure, i.e. $^{990}/_{25} \approx 40$ as presented in Table 3.7 (to be discussed in the next section), is comparable to the highest allophonic complexities tested so far in studies building upon Peperkamp et al.'s (2006) framework, viz. those attained by Martin et al.'s (2009) random-partitioning technique, i.e. $^{1800}/_{42} \approx 43$ on Japanese and $^{2200}/_{50} = 44$ on Dutch, as presented in Table 3.4. Although studies should be compared with all proper reservations because of differences in languages, corpora, and methodology, the experiments to be presented in subsequent chapters will therefore include evaluations of the robustness of our models to allophonic variation in a way comparable to Martin et al.'s study.

In order to guarantee the validity and the coherence of the allophonic inventories we derived, we performed a number of sanity checks on the various HMMs and GMMs computed by HTK. For each complexity, we successfully checked that:

— no two phonemes have identical phonemic HMMs;
— no two allophones have identical allophonic HMMs;
— no two allophones' GMMs have a null or almost null (viz. up to 0.1) acoustic dissimilarity, as will be defined in Chapter 4.

Table 3.7 — Derived allophonic inventories using HTK's decision-tree state-tying procedure. As for the size of the inventories, $\tilde{n}$ denotes the desired number of phones, and n the actual number of phones in each inventory. Values at the intersection of each column and row indicate the number of distinct allophones attested for that given phoneme at that given allophonic complexity. The penultimate column contains, for each phoneme, the number of distinct attested contexts. The last column contains phoneme frequencies, as observed in the CSJ.

$\tilde{n}$	50	100	150	200	250	300	350	400	450	500	550	600	650	700	750	800	850	900	950	1000	∞	*Freq.*
n	48	98	147	191	248	297	347	398	451	502	545	588	648	699	743	782	838	886	948	990	6539	
$n/\mathfrak{n}$	1.9	3.9	5.9	7.6	9.9	11.9	13.9	15.9	18.0	20.1	21.8	23.5	25.9	28.0	29.7	31.3	33.5	35.4	37.9	39.6	261.6	
/a/	10	15	28	42	58	63	72	72	82	98	103	108	130	153	153	167	167	183	201	202	585	*212554*
/e/	4	6	10	10	20	24	30	34	34	42	42	42	44	49	58	66	76	76	92	100	504	*109624*
/i/	4	6	12	12	15	24	32	40	42	48	61	62	73	89	92	91	82	82	91	91	532	*143285*
/o/	4	12	22	28	34	48	48	52	73	73	77	107	119	119	132	119	144	151	152	164	598	*165113*
/aː/	1	1	1	1	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	3	368	*6416*
/eː/	1	2	2	2	2	2	2	2	3	3	3	3	3	3	6	6	6	6	11	11	432	*23207*
/iː/	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	338	*4280*
/oː/	1	3	3	7	7	11	16	16	16	16	16	16	21	24	24	31	36	36	37	37	477	*37693*
/uː/	1	1	2	2	2	2	2	2	2	2	2	2	4	4	4	4	4	4	6	6	238	*16527*
/u/	1	4	4	3	4	6	6	8	8	9	11	11	11	11	13	13	13	13	15	14	501	*93155*
/w/	1	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2	2	2	2	4	61	*19925*
/j/	1	2	2	2	4	4	4	6	6	6	9	9	9	9	10	10	10	10	10	10	116	*30418*
/b/	1	1	1	1	1	1	1	1	1	1	2	2	2	2	2	3	3	3	3	3	131	*13248*
/d/	1	1	1	4	4	4	4	6	6	6	8	8	8	8	8	12	12	12	12	12	123	*47002*
/g/	1	1	1	2	2	2	2	2	2	2	3	3	3	3	4	5	5	5	5	5	124	*31778*
/p/	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	125	*6064*
/t/	3	15	18	21	24	29	35	39	39	48	48	48	48	52	54	63	71	71	71	72	153	*136392*
/k/	3	6	12	15	23	23	28	40	40	40	40	47	47	47	47	47	53	69	69	69	149	*105057*
/m/	1	3	6	9	9	9	9	12	12	12	12	12	12	12	12	15	15	20	24	24	112	*54567*
/n/	2	3	3	4	6	8	10	10	10	16	16	16	16	16	19	19	22	22	24	27	114	*91230*
/ɴ/	1	1	4	4	6	8	12	12	20	20	25	25	25	25	25	30	30	30	30	36	239	*51650*
/s/	1	9	9	16	16	16	16	24	35	40	40	40	40	40	47	47	55	55	55	55	136	*91532*
/z/	1	1	1	1	1	2	2	2	2	2	2	2	2	2	2	4	4	4	6	9	123	*20512*
/r/	1	1	1	1	4	4	9	9	9	9	9	9	15	15	15	15	15	19	19	24	123	*58354*
/h/	1	1	1	1	1	2	2	4	4	4	10	10	10	10	10	9	9	9	9	9	137	*22383*

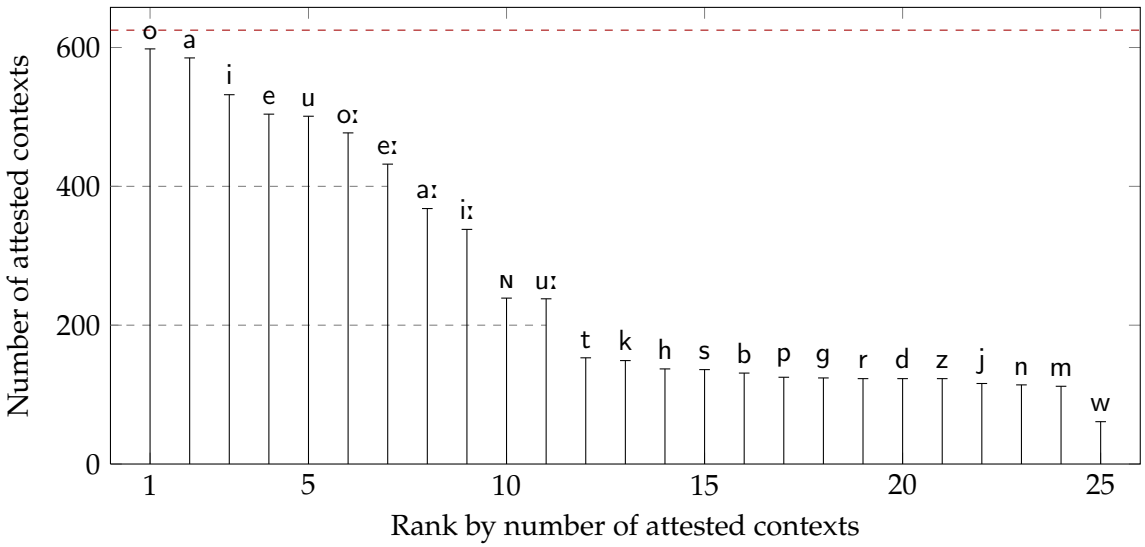


Figure 3.6 — Distribution of phonemes' number of attested contexts in the CSJ, as a function of ranks in the context table. The red line marks the theoretical maximum number of bilateral contexts a phoneme can have, i.e. $|\mathfrak{P}|^2 = 25^2 = 625$.

3.3 Allophonic inventories

In this section, we first examine the allophonic inventories we derived using HTK's decision tree-based state-tying procedure. We then go on to discuss how our data relate to aforementioned descriptions of allophonic processes in Japanese (Miller, 1967; Kinda-Ichi & Maës, 1978; Hinds, 1990; Ito & Mester, 1995; Okada, 1999; CSJ DVD, 2004; Vance, 2008; Wikipedia, 2012).

Upper bounds and expectations All experiments in computational linguistics and natural language processing are constrained by the properties (desirable or not) of their input data. In the case at hand, possible allophonic inventories are constrained by the empirical phonotactics of Japanese—that is to say, each phoneme's attested contexts in the CSJ. Furthemore, and because we assumed that there can be no more than one phone per phoneme per context (Assumption 3.2), we can easily compute the higher bound on the attainable number of phones given our corpus.

Following our assumption, the theoretical maximum number of allophones for each phoneme is given by $|\mathfrak{P}|^2 = \mathfrak{n}^2$, i.e. $\mathfrak{n}$ possible preceding contexts and $\mathfrak{n}$ possible following contexts, that is to say $25^2 = 625$ in the case at hand. As a consequence, the theoretical maximum number of phones is given by $|\mathfrak{P}|^3 = \mathfrak{n}^3$, i.e. one allophone per phoneme per context, that is to say $25^3 = 15625$. However, because of the inherent phonotactic constraints of the Japanese language and, perhaps, sampling limitations in the constitution of the CSJ, not every phoneme is attested in all possible contexts. Actually, in our data, there is not a single phoneme that is attested in all contexts. Indeed, as presented in Figure 3.6, the distribution of the number of attested contexts for each phoneme follows a distribution reminiscent of the frequency distribution reported in Figure 3.2: whereas few phonemes are attested in most contexts, most phonemes are attested in few contexts. The ratio of attested to possible contexts in the CSJ ranges from 598/625 ($\approx$ 96%) for /o/ down to 61/625 ($\approx$ 10%) for /w/. The empirical maximum number of phones is given by the sum, for each phoneme, of the number of attested contexts; altogether, only 6539 possible contexts are attested out the 15625 possible ones ($\approx$ 42%), a fact that quantifies the strentgh of phonotactic constraints in Japanese. Surprisingly, ordering all phonemes by their respective number of attested contexts brings, almost perfectly, broad natural classes to light: short vowels appear to be the least constrained phonemes, followed by long vowels and, finally, consonants and glides. It is worth noting that, however, there is no clear relation between a phoneme's frequency and the number of contexts it appears in, as illustrated by the scatter plot in Figure 3.7.

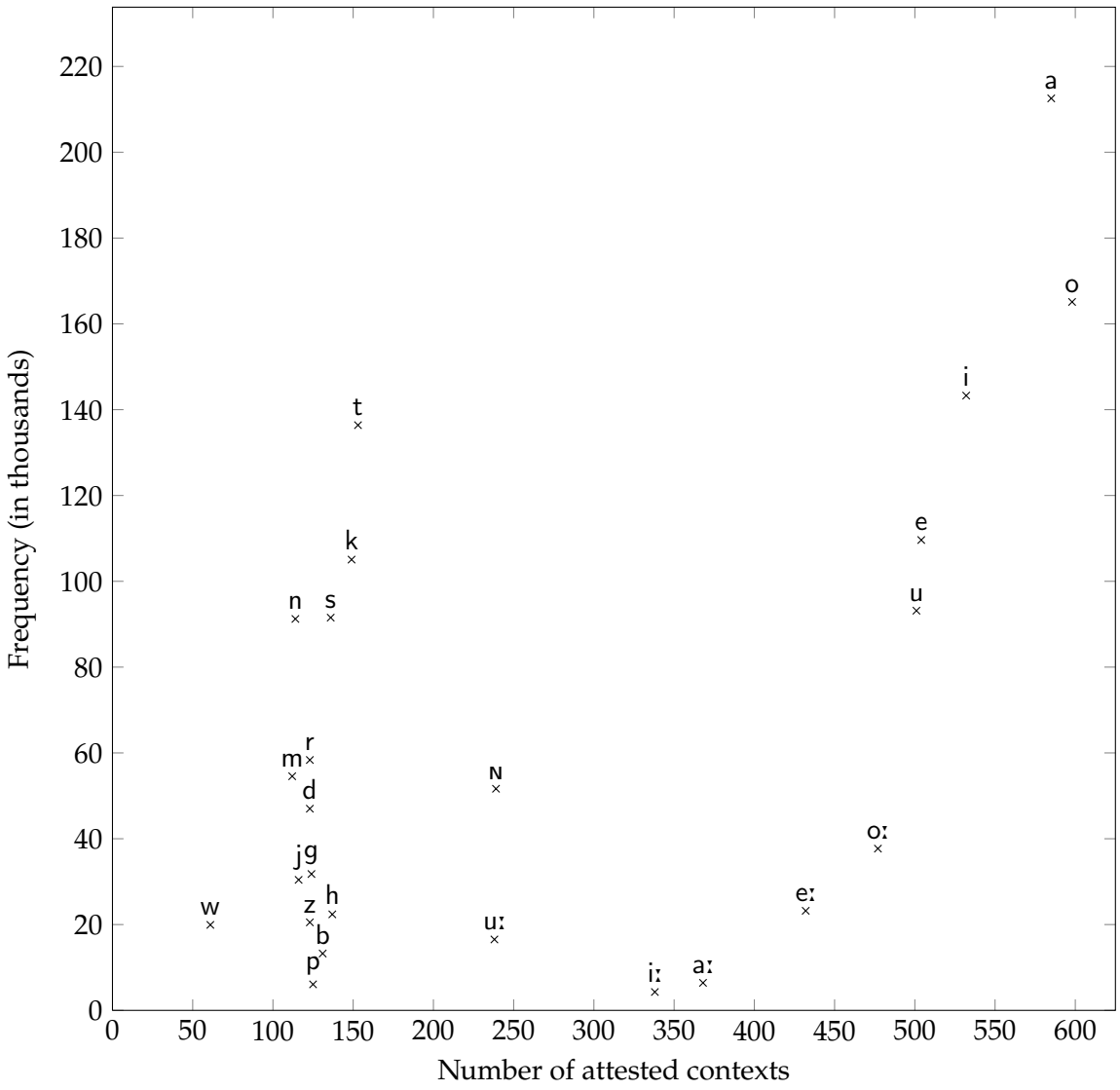


Figure 3.7 — Phoneme frequency as a function of the number of attested contexts in the CSJ.

3.3.1 Data-driven phonotactics

Further exploration of the phonotactics of Japanese is needed in order to know whether patterns emerge and, if so, if these patterns are linguistically coherent. To this end, we cross-tabulated every phoneme and possible context, checking whether the former is attested in the latter. The results are reported in the matrix presented in Table 3.8, where the symbols at the intersection of each column and row indicate whether the column phoneme is attested as a preceding or following context of the row phoneme, ignoring word boundaries for the purpose of this particular analysis.

First, it is worth pointing out that the observable diagonal symmetry in this matrix is not surprising: it is simply an illustration of our prior observation that every phoneme is an attested context of its attested contexts. The most noticeable pattern in this empirical view of the phonotactics of Japanese is the fact that absolutely all pure consonants (i.e. not the approximants) are attested as contexts of all vowels and, similarly, that all vowels are attested as contexts of all consonants. Moreover, the approximants /j/ and /w/ are only attested in vocalic contexts and, in the case of /j/, after a consonant. Furthermore, the black diagonal indicates that only stops, /b/, /d/, /g/, /p/, /t/, and /k/, fricatives, /s/, /z/, and /h/, as well as the moraic obstruent /ɴ/ (and not liquids or nasals) occur as geminated consonants. Although discussing each and every bullet in this matrix beyond the scope of the present study, previous observations correlate well with

Table 3.8 — An empirical view of the local phonotactics of Japanese. The symbols read as follows: ● indicates that the column phoneme is attested as both a preceding and a following context of the row phoneme, ◐ indicates that the column phoneme is only attested as a preceding context, ◑ indicates that the column phoneme is only attested as a following context, and ○ indicates that the column phoneme is not attested as a context of the row phoneme.

	a	e	i	o	u	aː	eː	iː	oː	uː	w	j	b	d	g	p	t	k	m	n	ɴ	s	z	r	h
a	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●
e	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●
i	●	●	●	●	●	●	●	●	●	●	●	◑	●	●	●	●	●	●	●	●	●	●	●	●	●
o	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●
u	●	●	●	●	●	●	●	●	●	●	◑	●	●	●	●	●	●	●	●	●	●	●	●	●	●
aː	●	●	●	●	●	●	●	●	●	◐	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●
eː	●	●	●	●	●	●	●	●	●	◐	●	◑	●	●	●	●	●	●	●	●	●	●	●	●	●
iː	●	●	●	●	●	●	●	●	●	◐	●	◑	●	●	●	●	●	●	●	●	●	●	●	●	●
oː	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●
uː	●	●	●	●	●	◑	◑	◑	●	●	◑	●	●	●	●	●	●	●	●	●	●	●	●	●	●
w	●	●	●	●	◐	●	●	●	●	◐	○	○	○	○	○	○	○	○	○	○	◐	○	○	○	○
j	●	●	◐	●	●	●	◐	◐	●	●	○	○	◐	◐	◐	◐	◐	◐	◐	◐	◐	○	○	◐	◐
b	●	●	●	●	●	●	●	●	●	●	○	◑	●	○	○	○	○	○	○	○	◐	○	○	○	○
d	●	●	●	●	●	●	●	●	●	●	○	◑	○	●	○	○	○	○	○	○	◐	○	○	○	○
g	●	●	●	●	●	●	●	●	●	●	○	◑	○	○	●	○	○	○	○	○	◐	○	○	○	○
p	●	●	●	●	●	●	●	●	●	●	○	◑	○	○	○	●	○	○	○	○	◐	○	○	○	○
t	●	●	●	●	●	●	●	●	●	●	○	◑	○	○	○	○	●	○	○	○	◐	○	○	○	○
k	●	●	●	●	●	●	●	●	●	●	○	◑	○	○	○	○	○	●	○	○	◐	○	○	○	○
m	●	●	●	●	●	●	●	●	●	●	○	◑	○	○	○	○	○	○	○	○	◐	○	○	○	○
n	●	●	●	●	●	●	●	●	●	●	○	◑	○	○	○	○	○	○	○	○	◐	○	○	○	○
ɴ	●	●	●	●	●	●	●	●	●	●	◑	◑	◑	◑	◑	◑	◑	◑	◑	◑	●	◑	◑	◑	◑
s	●	●	●	●	●	●	●	●	●	●	○	○	○	○	○	○	○	○	○	○	◐	●	○	○	○
z	●	●	●	●	●	●	●	●	●	●	○	○	○	○	○	○	○	○	○	○	◐	○	●	○	○
r	●	●	●	●	●	●	●	●	●	●	○	◑	○	○	○	○	○	○	○	○	◐	○	○	○	○
h	●	●	●	●	●	●	●	●	●	●	○	◑	○	○	○	○	○	○	○	○	◐	○	○	○	●

ubiquitous observations in most—if not all—aforementioned theoretical accounts of phonotactic constraints in Japanese: the elementary syllable (or mora) type in Japanese is CV (i.e. a consonant onset followed by a vowel nucleus), the approximant /j/ appears only in CjV-type syllables, and geminate consonants are limited to obstruents and the moraic nasal, respectively.

Yet another frequency effect All things considered, combining the statistics reported in Figure 3.2 (viz. the distribution of phoneme frequencies) and Figure 3.6 (the distribution of phonemes' number of attested contexts) to this data-driven examination of the phonotactic constraints at stake in the CSJ allows us to fully discuss the figures precedently reported in Table 3.7. As previously argued, allophonic variation is an inherently contextual process: in any language, each realization of a given phoneme is constrained by the surrounding phonemes. Therefore, one could expect that the more contexts a phoneme appears in, the more diverse its realizations would be. This conjecture is simply not true.

An example of this can be found considering the case of the long vowels /aː/, /eː/, /iː/, /oː/, and /uː/. As observed from the graph in Figure 3.6, these phonemes form the second best group in terms of the number of attested contexts: broadly speaking, each long vowel is attested in about twice as many contexts as any consonant. Similarly, the rows and columns devoted to the phonotactics of long vowels in Table 3.8 are mostly filled with black bullets, indicating that not only do long vowels occur in many contexts in Japanese, they also occur in very disparate

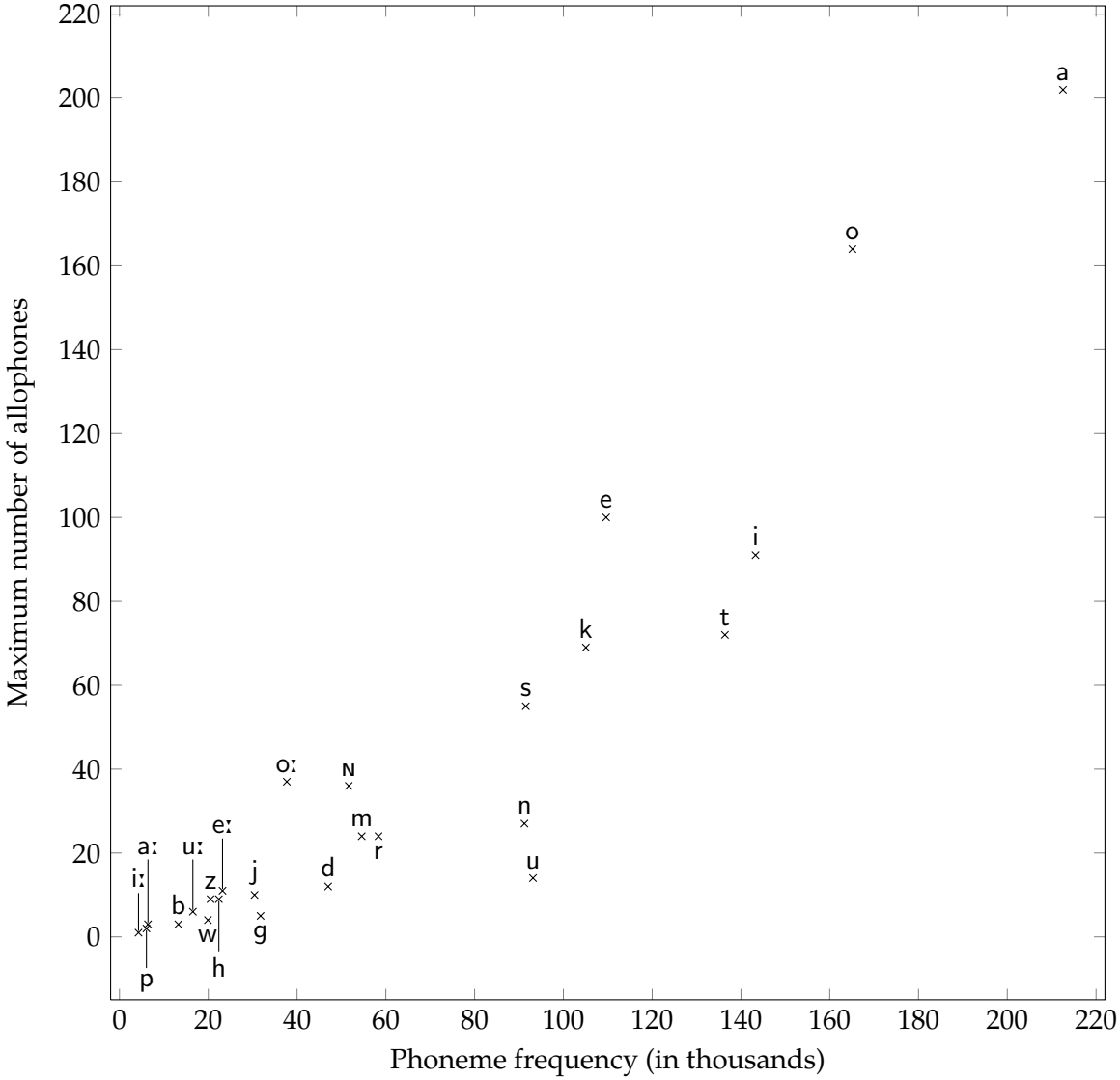


Figure 3.8 — Attested number of allophones at maximal allophonic complexity, i.e. $^{990}/_{25}$ allophones per phoneme, as a function of the frequency observed in the CSJ.

contexts. However, HTK's acoustics-based state-tying procedure yielded, at any allophonic complexity, a relatively low number of allophones for these phonemes. This contrast is best illustrated by also considering the situation for short vowels and at the maximal allophonic complexity considered in the present study, i.e. $^{990}/_{25}$ allophones per phoneme. Indeed, both classes are observed in very similar contexts in the CSJ; yet, whereas hundreds of allophones were defined by HTK for all but one short vowels (91 for /i/, 100 for /e/, 164 for /o/, and 202 for /a/), only a few dozen allophones were derived for long vowels altogether (1 for /iː/, 3 for /aː/, 6 for /uː/, 11 for /eː/, and 37 for /oː/). Most of all, even if /iː/ is attested 4280 times and in 338 distinct contexts in our corpus, HTK could not yield more than a single phone for this phoneme. Because the algorithm underlying HTK's state-tying procedure relies on the likelihood of candidate allophonic inventory given the acoustic information in the corpus, we speculate that the 4280 realizations of /iː/ are so homogeneous that splitting them into two or more subsets was never a statistically interesting solution, at least with regards to the global likelihood of the putative inventory.

By contrast, there is a strong, positive, and more than linear correlation between the number of allophones produced by HTK and phoneme frequencies, as illustrated in the scatter plot in Figure 3.8. Thus, in the interest of learning a phonemic inventory from speech data, more than

the number of attested phonemes, it appears to us that it is the frequency of a given phoneme that really dictates the variability in this phoneme's realizations. However, further research on this point would be needed, using other speech-processing techniques and corpora of different sizes and languages, to be able to determine whether or not this observation is specific to our particular setup.

3.3.2 O theory, where art thou?

In the final section of this chapter, we attempt to compare the output of HTK's acoustics-based state-tying procedure to allophonic rules presented in theoretical descriptions of allophonic processes in Japanese. However, it is worth immediately highlighting that a full qualitative (i.e. manual) examination of HTK's output is hardly feasible. The reason for this is twofold. First, whereas theoretical descriptions of allophony include rules in which all elements (i.e. the target phoneme, the application context, and the resulting phone) are specified, e.g. h → ɸ / — u using Chomsky & Halle's (1968, p. 332) notation, HTK only instances the target phoneme and the application contexts, the resulting allophone being represented by the HMM of its actual realizations in the training data. Therefore, in HTK's output, no phonetic symbol is (nor can be) assigned to allophones, and their definition is limited to the set of all contexts in which they are realized. Second, and most importantly, the procedure we described in the preceding section generates one decision tree for every phoneme, HMM emitting state, and allophonic complexity—that is to say $25 \times 3 \times 20 = 1500$ different decision trees in the case at hand.

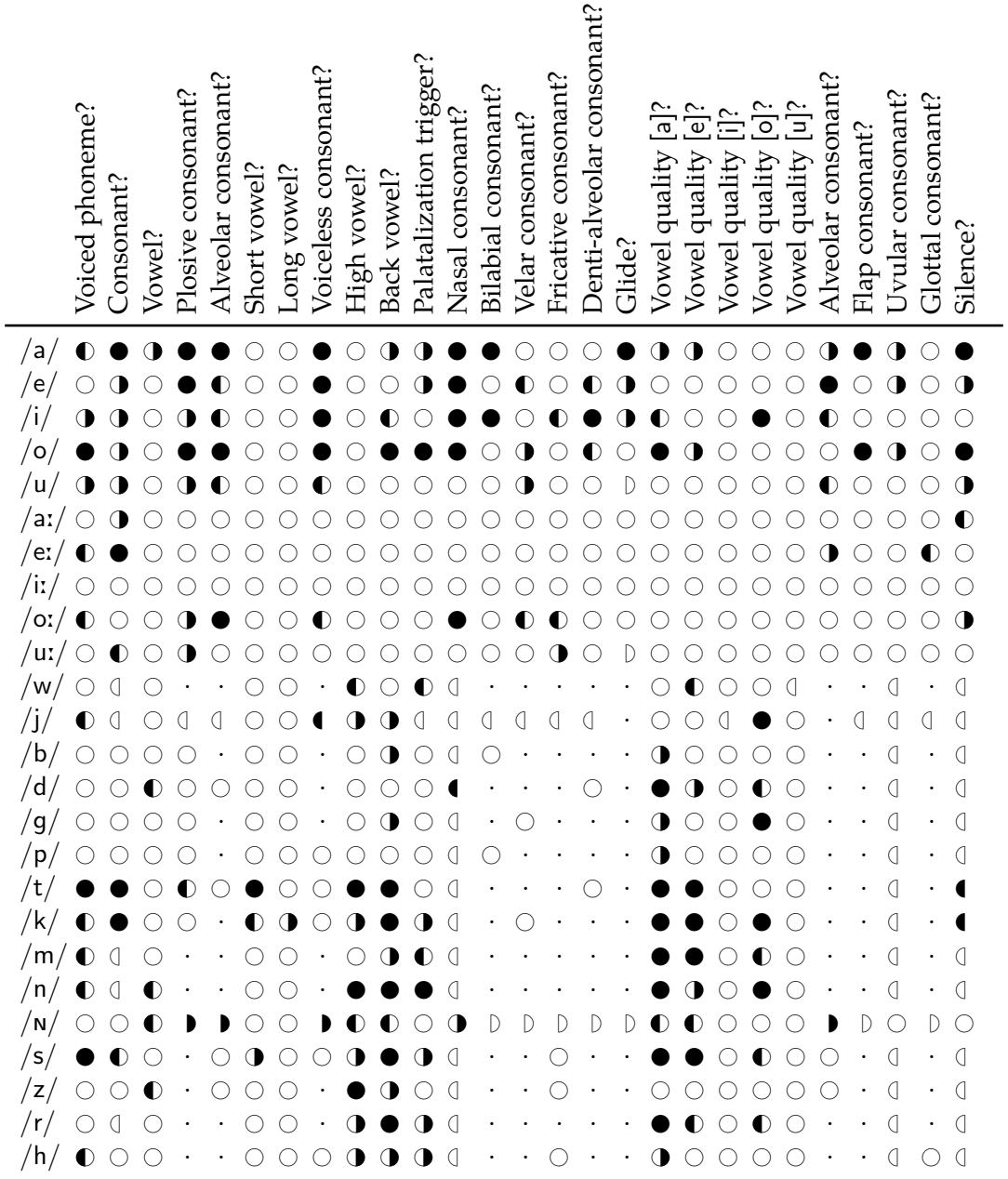
Established allophonic rules In Japanese, various allophonic rules constrain the realizations of both consonants and vowels; we will review the most commonly discussed processes.

As previously mentioned in this chapter, consonants are subject to a regular palatalization process when followed by the customary palatalization triggers /i/ and /j/, e.g. s → ʃ / — i (Miller, 1967) and h → ç / — i (Miller, 1967; Ito & Mester, 1995; Wikipedia, 2012). The vowel /u/, too, affects the consonants it follows by the application of allophonic rules such as h → ɸ / — u (Miller, 1967; Ito & Mester, 1995; Wikipedia, 2012) and t → ʦ / — u (Miller, 1967); this explains why, for example, the underlying form in Japanese of the English loanword *tsunami* is actually /tunami/. The process of lenition (a.k.a. weakening, whereby a consonant becomes more vowel-like) is also attested in Japanese for /b/ and /g/ when they occur in an intervocalic context, as examplified by the rules b → β / V — V and g → ɣ / V — V (Kinda-Ichi & Maës, 1978; Wikipedia, 2012). Finally, most if not all aforementioned sources state that the moraic nasal phoneme /ɴ/ is subject to strong assimilation processes: it is said to be bilabial before bilabial consonants, i.e. ɴ → m / — {p, b, m}, coronal before coronals, i.e. ɴ → n / — {t, d, n, r}, velar before velars, i.e. ɴ → ŋ / — {k, g}, and uvular [ɴ] in other contexts; put another way, this phoneme only carries information about nasality and obstruction, its place of articulation depending on that of the following phoneme.

As far as vowels are concerned, the major allophonic process in Japanese is devoicing: short vowels, and especially the high vowels /i/ and /u/, become devoiced between two voiceless consonants (Kinda-Ichi & Maës, 1978; Wikipedia, 2012), e.g. i → i̥ / C̥ — C̥ and u → u̥ / C̥ — C̥.

State-tying's aftermath We now go on to examine whether these allophonic rules can be found in HTK's output. Before turning to the actual decision trees, the first step in our analysis concerns the linguistic questions we provided to HTK, i.e. that we are interested in reviewing, for every phoneme in the inventory, the questions that proved to be statistically relevant for an acoustics-based definition of its allophones. To this end, we cross-tabulated all questions and phonemes, distinguishing between questions that were actually used by HTK, questions that were not used though they define attested contexts for the phoneme at hand, and questions that could not be used because of phonotactic constraints on that phoneme. The results are reported in the matrix presented in Table 3.9.

Table 3.9 — Actual usage of the questions presented in Table 3.6 by HTK's decision tree-based state-tying procedure, across all allophonic complexities. The symbols read as follows: the left half denotes the phoneme's preceding contexts, the right half denotes the following contexts, a filled half-circle marks a question used at least once in the definition of that phoneme's allophones, an empty half-circle marks a question that was never used though that phoneme occurs in at least one context of that question, a missing half-symbol marks that no context of that question is attested for that phoneme, and a dot marks a question whose contexts are never attested for that phoneme, neither before nor after.



Interestingly, patterns do emerge in this matrix. First and foremost, one can observe that both palatalization-related questions, i.e. *Palatalization trigger?* and *High vowel?*, were effectively used by HTK to define the following context of at least one allophone for most consonants, viz. /t/, /k/, /n/, /s/, /z/, /r/, and /h/. Hence, this confirms the major role of palatalization in Japanese phonology, be it truly allophonic or a mere consequence of our decision to represent *yōon* as Cj-type bigrams. Another, less suprising, pattern shows that vowels' realizations appear to be constrained by consonant contexts, and consonants' realizations by vowel contexts; an

observation once again coherent with the preponderance of CV syllables in Japanese: if CV is the major syllable type, then words and utterances may likely consist in a binary alternation of consonants and vowels. Finally, it is worth noting that the *Voiceless consonant?* question was used to define at least one allophone for each and every short vowel and that, for all short vowels but /u/ (the least frequent of all), this very question was used by HTK as both a preceding context and a following context; we can thus reasonably speculate that these results corroborate linguists' statement about the frequent devoicing of short vowels between two voiceless consonants.

Unfortunately, no other known property of Japanese seems to be represented in Table 3.9: no intervocalic context emerged in the descriptions of the allophones of both /b/ and /g/, /u/ does not appear to constrain the realizations of /h/, /t/ or any other phoneme, and the realizations of /ɴ/ are conditioned by as many preceding as following contexts. However, Table 3.9 can not provide a full examination of HTK's output: because linguistic questions are applied sequentially from the most to the least general, it is well possible that, for example, the affricate realization [ʦ] of /t/ before /u/ was taken into account by the early *Voiced phoneme?* or *Back vowel?* questions, rather than the late *Vowel quality [u]?* question.

Allophonic trees In order to gain further insights on the allophones derived by HTK, we now go on to explore the decision trees themselves. As previously stated, 1500 trees were generated and, thus, providing a full examination of every single tree appears to be intractable, if relevant at all. Therefore, we concentrated our effort on phonemes whose decision trees at a given allophonic complexity are true refinements of their decision trees at lower complexities, what we refer to as allophonic trees. More precisely, because allophones are represented as context sets by HTK, an allophonic tree guarantees the following property: each allophone at a given allophonic complexity has to be accounted for at the immediately higher complexity either by the exact same context set, or by the union of this context set and at least one other set.

It is worth noting that because each state-tying procedure is performed independently of the others, increasing the desired allophonic complexity does not necessarily increase the number of allophones computed for every phoneme. Indeed, as reported in Table 3.7, the number of allophones of a given phoneme can even decrease as the global allophonic complexity increases; in the case at hand, this property is attested for /u/ at $^{150}/_{25}$ and $^{950}/_{25}$ allophones per phoneme, for /i/ at $^{750}/_{25}$ and $^{800}/_{25}$, and for /o/ and /h/ at $^{750}/_{25}$. Because their respective number of allophones is neither always constant (i.e. the partition is identical) nor increasing (i.e. the partition is refined) as the global allophonic complexity increases, the phonemes /i/, /u/, /o/, and /h/ can obviously not be represented as allophonic trees.

Notwithstanding this rather technical limitation, it turns out that allophonic trees are observed only for 7 phonemes, viz. /eː/, /iː/, /j/, /w/, /g/, /p/, and /s/. Moreover, it is worth noting that the case of /iː/ is trivial, as this phoneme always received a single allophone, even at $^{990}/_{25}$ allophones per phoneme; hence, its allophonic tree is reduced to the root. The resulting, non-degenerate allophonic trees observed for /p/, /w/, /g/, /j/, /eː/, and /s/ (arranged by increasing maximum allophonic complexity) are presented in Figures 3.9, 3.10, 3.13, 3.11, 3.12, and 3.14, respectively. By convention, decision trees read as follows:

— the left branch of a node denotes a negative answer to the linguistic question in the node, and the right branch denotes a positive answer (Young et al., 2006; p. 145);
— L and R stand for the preceding (i.e. left) and the following (right) contexts, respectively;
— $\mathrm{p}_{x\#y}$ denotes the x-th state of the y-th allophonic HMM modeling the realizations of the phoneme p;
— tied HMM states, e.g. $\mathrm{s}_{4\#1}$ and $\mathrm{s}_{4\#5}$ in Figure 3.14, are marked by a dashed circle.

The allophonic tree derived from the realizations of /p/, presented in Figure 3.9, confirms our previous observation that frequency plays a major role in HTK's behavior: whereas this phoneme is potential target for both allophonic and *yōon*-related palatalizations (CSJ DVD, 2004), the only context used by HTK to define its allophones is {a, aː}, i.e. the most frequent vowel quality in our corpus. The allophonic tree for /w/, too, is surprising; as presented in Figure 3.10,

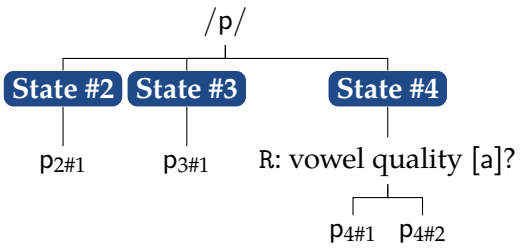


Figure 3.9 — Allophonic tree for the 2 allophones of /p/, as computed by HTK's decision tree-based state-tying procedure.

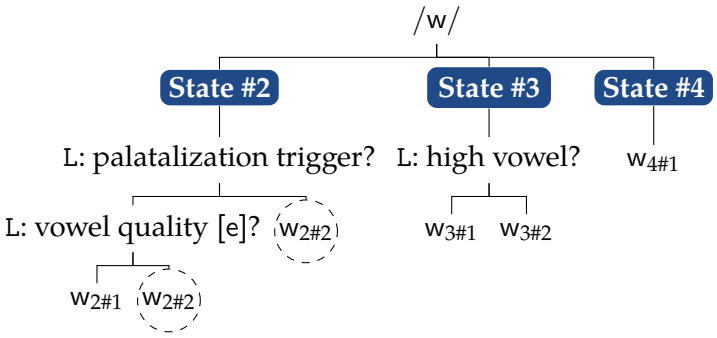


Figure 3.10 — Allophonic tree for the 4 allophones of /w/, as computed by HTK's decision tree-based state-tying procedure.

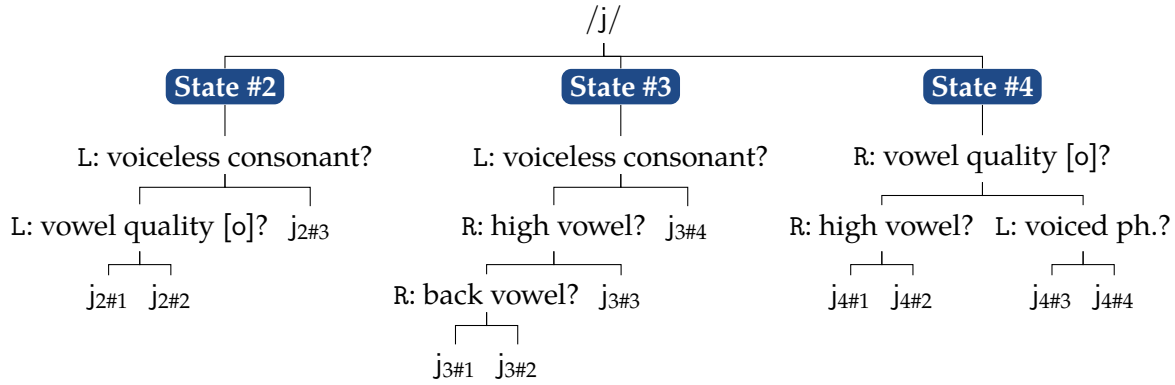


Figure 3.11 — Allophonic tree for the 10 allophones of /j/, as computed by HTK's decision tree-based state-tying procedure.

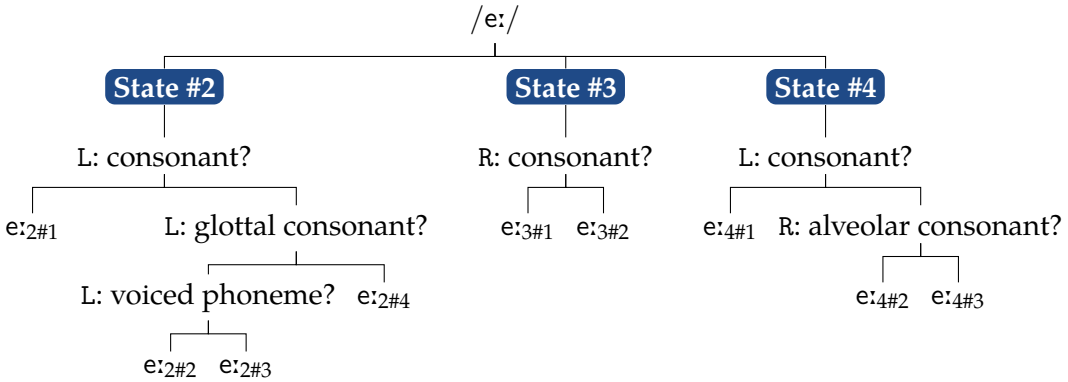


Figure 3.12 — Allophonic tree for the 11 allophones of /eː/, as computed by HTK's decision tree-based state-tying procedure.

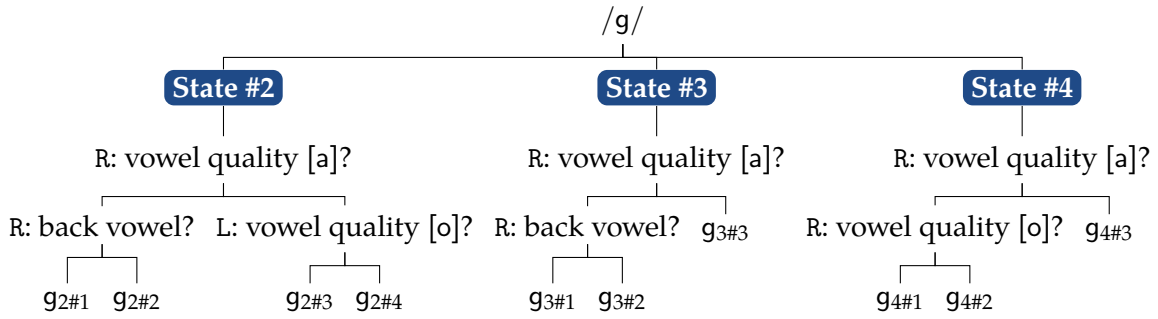


Figure 3.13 — Allophonic tree for the 5 allophones of /g/, as computed by HTK's decision tree-based state-tying procedure.

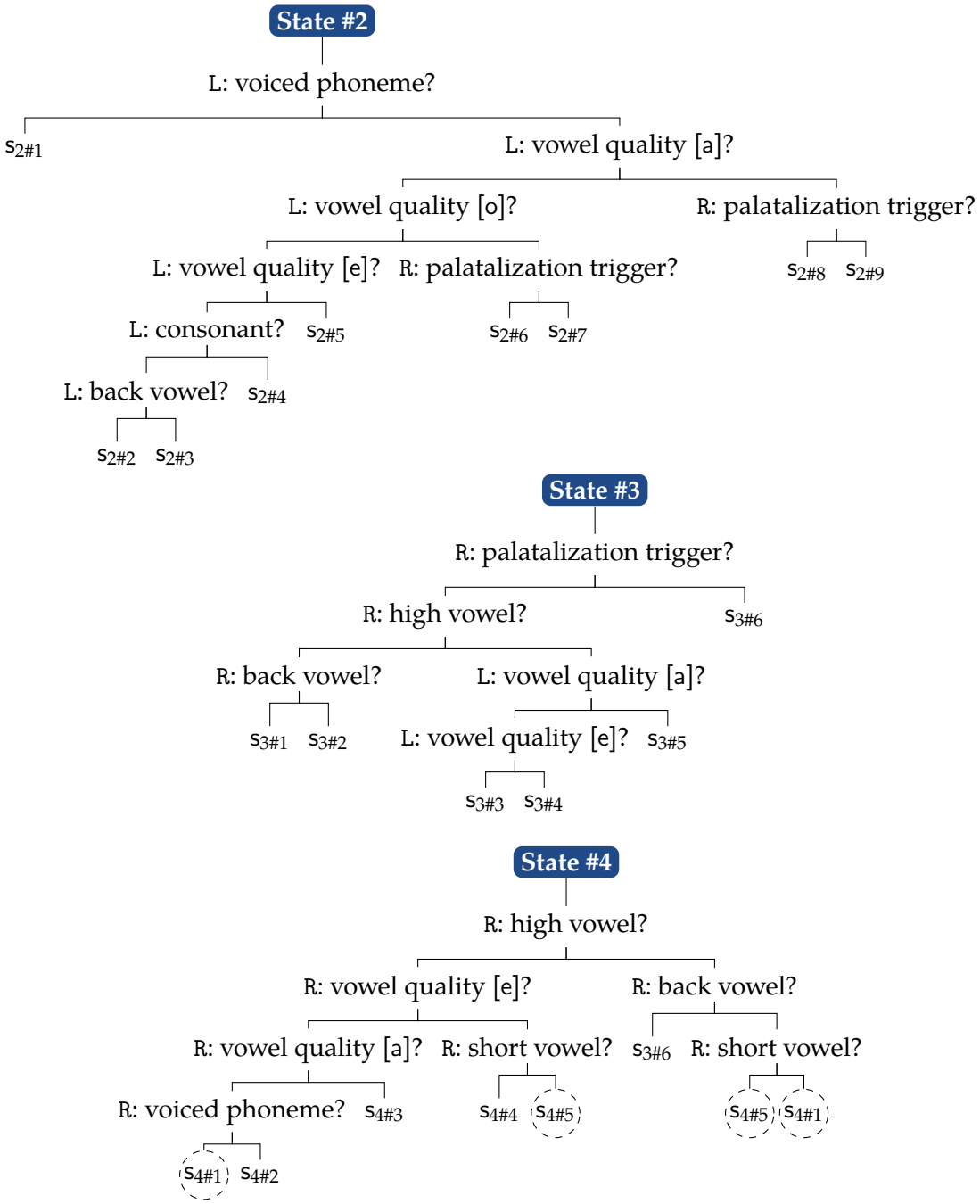


Figure 3.14 — Allophonic tree for the 55 allophones of /s/, as computed by HTK's decision tree-based state-tying procedure.

this phoneme's realizations appear to be most constrained by preceding high or front vowels, a fact that, to our knowledge, has never been discussed in the literature. Similarly, the allophonic processes underlying the realizations of the phoneme /j/ are unaccounted for; indeed, the voiceness of the preceding phoneme and the height of the following vowel appear to be the most relevant contexts to model the different realizations of this phoneme, as attested by the repeated usage of the questions *Voiceless consonant?*, *Voiced phoneme?*, and *High vowel?* in the decision tree presented in Figure 3.11. By contrast, the allophonic tree derived the long vowel /eː/ may denote a known phonemenon, though most commonly observed for short vowels, viz. vowel devoicing between voiceless consonants; as presented in Figure 3.12, most context sets comprise or consist in voiceless consonants as examplified by the use of the questions, *Glottal consonant?*, *Consonant?* and, to a lesser extent and by the negative, *Voiced phoneme?*. Although the realizations of /g/ appear to be solely constrained by vowels, 6 out of 7 questions relate to the following context only, as presented in Figure 3.13; further examination of the recordings of the CSJ would be needed to determine whether the intervocalic lenition rule g → ɣ / V — V is attested for this phoneme in the dataset. Gaining sense of all processes at stake in an allophonic tree as considerable as the one derived for /s/ is not straightforward; however, a major feature of the phonology of Japanese seems to be corroborated by HTK's output: the highest-level questions defining the following context of this phoneme's realizations are *Palatalization trigger?* and *High vowel?*, once again confirming the ubiquity of palatalization processes in Japanese.

Data-driven linguistics All in all, although some properties of the allophonic inventories that were derived using HTK's acoustics-based state-tying do relate to theoretical accounts of the phonology of Japanese, not all allophonic rules could be retrieved from the decision trees.

In our opinion, such observations do not hinder the plausibility of using allophones à la HTK to model early language acquisition. The reason for this is threefold. First, whereas the primary source of data in HTK's procedure is the audio recordings, most—if not all—aforementioned grammars of the phonology of Japanese focus on articulatory descriptions of allophonic rules, relegating fine-grained acoustic considerations to the background (a notable exception being Vance, 2008). Second, we have shown that the computation of our acoustics-based allophones is heavily influenced by phoneme frequencies; thence, whereas it is surely appropriate to include a rule such as h → ç / — i in an allophonic grammar of Japanese, it is unlikely that a low-frequency phoneme such as /h/ will have a significant weight during a global partitioning of all phones in the CSJ. Finally, speech is the major medium by which infants receive linguistic input and, if our allophones are acoustically different enough to be clustered in such a way by HTK, then it is likely that they are perceptually distinguishable by infants.

CHAPTER 4
INDICATORS OF ALLOPHONY

Studies building upon Peperkamp et al.'s original experiments have defined various numerical criteria that—tracking statistical regularities in a phonetically transcribed corpus—have been used to discover which phones are actually allophones of some (yet to be discovered) phoneme in the language at hand. For instance, these criteria include measures of near-complementary distributions (Peperkamp et al., 2006; Le Calvez, 2007; Le Calvez et al., 2007; Dautriche, 2009; Martin et al., 2009; Boruta, 2009, 2011b), articulatory or acoustic similarity (Peperkamp et al., 2006; Le Calvez, 2007; Le Calvez et al., 2007; Dautriche, 2009), and, more recently, lexical dissimilarity (Martin et al., 2009; Boruta, 2009, 2011b). However, no common framework has yet been proposed to systematically define and evaluate these data-driven measures, and Peperkamp et al.'s (2006) original, distributional learner has been extended with so called constraints, filters, and cues. In this chapter, we show that all aforementioned measures indicating, to some extent, which phones are allophones can actually be redefined as various instances of a single, cohesive concept, viz. phone-to-phone measures of dissimilarity referred to as *indicators of allophony*.

This chapter is divided into six main sections. In Section 4.1, we define our goals as well as the formal data structures that will be used throughout this study. Acoustic, temporal, distributional and lexical indicators of allophony are then precisely defined in Section 4.2. Section 4.3 contains a discussion of the various standardization techniques we used so that a consistent examination of indicators' performance could be carried out. In Section 4.4, we present a first assessment of indicators' informativeness in terms of prognoses of allophony, i.e. preliminary inspections of the correlation between our indicators' value and the allophony relation. Various classification experiments are then reported and discussed in Section 4.5. Finally, Section 4.6 contains a general assessment of the work presented in this chapter.

4.1 Allophony: definitions and objectives

In this section, we first define the concept of indicator of allophony, as well as the mathematical objects we will use to represent and manipulate such indicators. We then go on to specify the objectives of our artificial learner, viz. predicting whether or not two phones are allophones.

4.1.1 Phones, phone pairs, and allophones

As discussed in Chapter 2, the phonology of every natural language revolves around its own inventory of phonemes, i.e. a finite (and rather small) set of abstract sound categories. In turn, every phoneme can be realized by a (finite, following Assumption 3.1) number of different phones, depending on its context. Indeed, as summarized by Lyons (1968; pp. 100–101):

> "the distinction between [two phones] never has the function of keeping apart different words [...]; it is not a *functional* difference: it is a phonetic difference, but not a phonological,

> or phonemic, difference [...]. In cases of this kind we say that the phonetically distinguishable pairs of speech sounds are positional variants, or *allophones*, of the same phoneme.❞

Allophony can thus be thought of as a polyadic relation whereby allophones consist in sets of phones brought together on the grounds that they all are realizations of the same, underlying phoneme. To draw a non-technical analogy, the allophony relation shares many formal properties with the siblings relationship: a sibling (an allophone) is one of various inviduals (various phones) having a parent in common (having the same underlying representation). Moreover, just as the siblings relationship may involve two or more individuals, the allophony relation may involve two or more phones. Indeed, we observed in Chapter 3 that not all phonemes have the same number of allophones (cf. Table 3.7). In fact, observing that two or more phonemes have the same number of allophones would be nothing but a mere statistical accident.

Despite this inherent property of natural languages, Peperkamp et al. (2006) introduced a strictly pairwise computational framework whereby, for any phoneme, allophones are discovered two by two. For instance, let $\{p_1, p_2, p_3\} \subseteq P$ denote the phones accounting for the contextual realizations of an hypothetical phoneme $\mathfrak{p} \in \mathfrak{P}$. In Peperkamp et al.'s framework, learning the allophony relation between these three phones amounts to learning that all three pairs $\{p_1, p_2\}$, $\{p_1, p_3\}$, and $\{p_2, p_3\}$ are each made up of two allophones. Let $n \equiv |P|$ denote the size of a given allophonic inventory P; put another way, the task of Peperkamp et al.'s artificial learner consists in deciding for each of the $n(n-1)/2$ possible pairs of phones whether or not the two phones it comprises are allophones. It is worth pointing out that an additional step would be required to infer the phonemic inventory of the language at hand: detecting which pairs of phones are allophones is indeed insufficient to discover polyadic sets of allophones. In other words, the fact that p_1 belongs to both $\{p_1, p_2\}$ and $\{p_1, p_3\}$ needs to be accounted for. In this chapter, the focus in on Peperkamp et al.'s original pairwise allophony task. The limitations of this pairwise approach will be further discussed in Chapter 5 and circumvented in Chapter 6.

4.1.2 Objectives: predicting allophony

The overall goal of Peperkamp et al.'s (2006) artificial learner is to reduce the inherent variability in speech, represented in the present study as a set of n phones $P \equiv \{p_1, \ldots, p_n\}$, to a finite set of abstract sound categories, i.e. the $\mathfrak{n}$ phonemes $\mathfrak{P} \equiv \{\mathfrak{p}_1, \ldots, \mathfrak{p}_\mathfrak{n}\}$ of the target language. Let us mention that we make the trivial assumption that $n > \mathfrak{n}$ throughout the present study, i.e. that there are strictly more phones than phonemes and, hence, that the problem is not already solved.

Following the formal assumptions we made in Section 3.2, every phone is the realization of one (and only one) phoneme, and every phoneme has at least one realization. Formally, we can thus define the set of phonemes $\mathfrak{P}$ as a partition of the set of phones P. More precisely, $\mathfrak{P}$ is a set of nonempty subsets of P such that every phone $p_i \in P$ belongs to exactly one of these subsets, i.e. the subset-like phonemes are both collectively exhaustive and mutually exclusive:

$$\bigcup \mathfrak{p}_h = P \qquad \text{and} \qquad \mathfrak{p}_h \cap \mathfrak{p}_{h'} = \varnothing \text{ if } h \neq h', \qquad h \in (1, 2, \ldots, \mathfrak{n}). \tag{4.1}$$

Hence, under this representation of phonemes as sets, it is worth noting that the number of allophones of a given phoneme $\mathfrak{p}_h$ is merely given by its cardinality $|\mathfrak{p}_h| \geq 1$.

A pairwise framework Formalizing the task introduced by Peperkamp et al. (2006), our goal is to learn an $n \times n$ symmetric Boolean matrix of allophony $\mathbf{A} \equiv [a_{ij}]$ where $a_{ij} \in \mathbb{B}$ denotes what will be referred to as the *allophonic status* of the pair of phones $\{p_i, p_j\} \subseteq P$. If $a_{ij} = 1$, the pair $\{p_i, p_j\}$ is said to be *allophonic* (i.e. p_i and p_j are allophones) and, otherwise, it is said to be *non-allophonic* (p_i and p_j are not allophones). The true, reference allophonic statuses can be derived from the phonemic partition $\mathfrak{P}$ of P into a reference matrix of allophony $\mathbf{A}^* \equiv [a^*_{ij}]$ whose values are given by

$$a^*_{ij} \equiv [\![\exists\, h \in (1, 2, \ldots, \mathfrak{n}), \{p_i, p_j\} \subseteq \mathfrak{p}_h]\!] \tag{4.2}$$

where $[\![\cdot]\!]$, the evaluation function, denotes a number that is 1 if the condition within the double brackets is satisfied, and 0 otherwise.

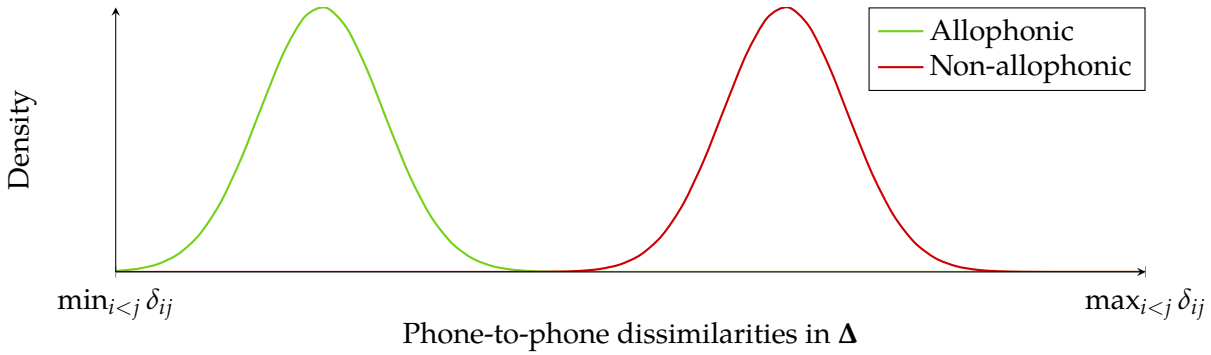


Figure 4.1 — Schematic representation of the distribution of an ideal indicator of allophony.

Indicators of allophony Following Peperkamp et al., we aim at predicting the allophonic status of any given phone pair in a corpus. In order to do so, the first step consists in gathering information about various cues and patterns in the input. In the present study, such empirical information will take the form of quantitative, phone-to-phone dissimilarity measurements that we will refer to as *indicators of allophony*, as defined in a previous study (Boruta, 2011b; p. 1):

> ❝Discrimination relies on indicators, i.e. linguistic properties which are correlated with allophony.❞

The fundamental hypothesis behind indicators of allophony is the aforementioned compactness hypothesis (Duin, 1999; Pękalska et al., 2003), i.e. the hypothesis that underlyingly similar objects are also close in their surface representations. Concretely, we assume that, for any given indicator, the likelihood of two phones being allophones is inversely correlated with their dissimilarity. Put another way, and as illustrated in Figure 4.1, we assume that:

Assumption 4.1 The values of an effective indicator of allophony follow a bimodal distribution whose modes emblematize the dichotomy between allophonic pairs and non-allophonic pairs, keeping the ones apart from the others.

Under this (implicit) assumption of a bimodal distribution, prior studies relied on the corollary that the two target statuses (i.e. allophonic and non-allophonic) can be materialized by searching for a threshold value at and above which phone pairs are classified as non-allophonic (Peperkamp et al., 2006; Le Calvez, 2007; Le Calvez et al., 2007; Boruta, 2011b). Formally, an indicator of allophony is represented as a square $n \times n$ dissimilarity matrix $\mathbf{\Delta} \equiv [\delta_{ij}]$ where δ_{ij} denotes the dissimilarity between the phones p_i and p_j, as estimated from observed data.

Allophony is, by definition, a symmetric relation, i.e. $\forall \{p_i, p_j\} \subseteq P, a^*_{ij} \Rightarrow a^*_{ji}$. Therefore, for the sake of simplicity, we require that the dissimilarity measures given by any indicator be symmetric, too, i.e. $\forall \{p_i, p_j\} \subseteq P, \delta_{ij} = \delta_{ji}$. Moreover, because of technical reasons to be discussed in Chapter 6, we also require that every dissimilarity matrix $\mathbf{\Delta}$ be non-negative, i.e. $\forall \{p_i, p_j\} \subseteq P, \delta_{ij} \geq 0$, and hollow, i.e. $\forall p_i \subseteq P, \delta_{ii} = 0$. Because of our assumption that the likelihood of two phones being allophones is inversely correlated with their non-negative dissimilarity, this last constraint entails that allophony is, per our definitions, a reflexive relation, i.e. $\forall p_i \subseteq P, a^*_{ii} = 1$. Although this appears to be inconsistent with our prior definitions of allophones as the *distinct* realizations of a given phoneme—returning to our previous analogy, one can not be one's own sibling—it has virtually no consequence in the case at hand. Indeed, whereas the elementary object in Peperkamp et al.'s framework is a pair of phones $\{p_i\} \cup \{p_j\} = \{p_i, p_j\}$, set objects can not include duplicate elements, i.e. $\{p_i\} \cup \{p_i\} = \{p_i\}$. Consequently, off-diagonal values are the only relevant values in any dissimilarity matrix $\mathbf{\Delta}$.

In the next section, each indicator is to be defined by its generator function $f : P \times P \to \mathbb{R}^+$ whereby the values of its dissimilarity matrix $\mathbf{\Delta}$ are given by $\delta_{ij} = f(p_i, p_j)$.

4.2 Building indicators of allophony

Before turning to the actual definitions, let us revisit the nature and the form of the data that are available to us in order to build indicators of allophony. On one hand, we can rely on acoustics-based allophonic HMMs modeling the realizations of every single phone in the CSJ and, on the other hand, we can monitor statistical patterns in a phonetic transcription of the CSJ where data have been coded at three different linguistic levels: phones, words, and utterances (cf. Chapter 3).

For the sake of brevity, indicators of allophony are insignificantly defined in this section as either similarity or dissimilarity measures; standardization issues will be discussed in Section 4.3. Furthermore, it is worth mentioning that each indicator's name includes a prefix blackboard-bold letter reminiscent of the type of information this indicator relies on: $\mathbb{A}$ for acoustic indicators, $\mathbb{T}$ for temporal indicators, $\mathbb{D}$ for distributional indicators, and $\mathbb{L}$ for lexical indicators.

4.2.1 Acoustic indicators

In this section, we describe the computation of two indicators relying on measures of the acoustic dissimilarity between phones. Although the relevance of phonetic criteria for the definition of phonemes has been discussed and, enventually, minimized in favor of an emphasis on distributional and functional criteria (Trubetzkoy, 1939; Austin, 1957), we reiterate our argument that disregarding actual sounds in a study of sound categories would appear to be contrieved. Moreover, and as previously stated, speech is the major medium by which infants receive linguistic input. We therefore consider acoustic information to be a legitimate cue to keep track of.

Prior scholarship Using the data at hand, the act of measuring the acoustic dissimilarity between two phones amounts to measuring the dissimilarity between the two allophonic HMMs that were trained to model their actual realizations in the recordings of the CSJ. We will thus focus in the present study on the measure of acoustic dissimilarity described by Dautriche (2009; implementation by Dupoux & Schatz, priv. comm.), for the sake of comparability—even though various seminal acoustic distance measures have been defined (e.g. Gray & Markel, 1976; Mermelstein, 1976; Everson & Penzhorn, 1988; Mak & Barnard, 1996). For this particular measure as in Section 3.2.2, it is worth noting that the following protocol is not ours; it is only presented here in order to ensure the reproducibility of the results to be presented in subsequent sections.

Concretely, Dautriche (2009) defined the acoustic dissimilarity between two phones $\{p_i, p_j\} \subseteq P$ as the dissimilarity between their respective HMMs. In that study, the latter dissimilarity is given by the dynamic time warping between the two sequences $\mathbf{s}_i$ and $\mathbf{s}_j$ made up of their respective emitting states, and using a symmetrization of the Kullback–Leibler divergence between the states' output probability density functions as the local distance function. The generator function of this indicator, which will be referred to as $\mathbb{A}$-DTW, is given by

$$\mathbb{A}\text{-DTW}(p_i, p_j) \equiv \text{DTW}(\mathbf{s}_i, \mathbf{s}_j; \text{SKLD}) \tag{4.3}$$

where SKLD denotes Dautriche's symmetrization of the Kullback–Leibler divergence (to be precisely defined hereafter), and $\text{DTW}(\mathbf{x}, \mathbf{y}; f)$ denotes the value of the optimal symmetric dynamic time warping that completely maps the sequences $\mathbf{x}$ and $\mathbf{y}$ onto one another using f as the local distance (Sakoe & Chiba, 1978; Giorgino, 2009). Here, dynamic time warping is used to find the optimal path among all paths from both HMMs' first emitting states to both HMMs' last emitting states. The search space of all such possible paths is illustrated in the lattice-like structure in Figure 4.2.

Dautriche's symmetrized Kullback–Leibler divergence is a straightforward extension of the original Kullback–Leibler divergence (henceforth KLD; Kullback & Leibler, 1951), a seminal measure of dissimilarity between two probability density functions. For two arbitrary density functions $f(x)$ and $f'(x)$, their KLD is given by

$$\text{KLD}(f \parallel f') \equiv \int f(x) \log \frac{f(x)}{f'(x)} \, \mathrm{d}x \tag{4.4}$$

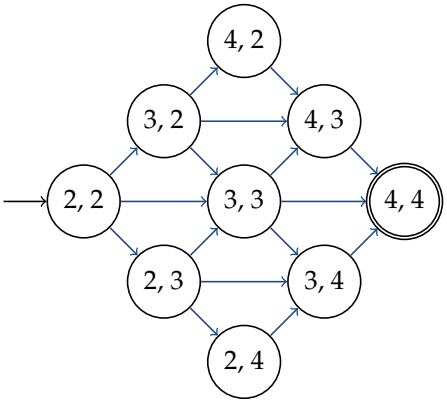


Figure 4.2 — Search space for the dynamic time warping between two HMMs' emitting states, as used in Dautriche's (2009) acoustic indicator. In node labels, ordered pairs of the form (x, y) denote the pair formed by the x-th state of the first HMM and the y-th state of the second.

and, for two b-variate Gaussian density functions g and g' as defined in Section 3.2.2, KLD has the following closed formed expression (e.g. Hershey & Olsen, 2007), using the notational shorthand $g \equiv g(\mathbf{x} \mid \theta)$ and $g' \equiv g(\mathbf{x} \mid \theta')$:

$$\mathrm{KLD}(g \,\|\, g') \equiv \frac{1}{2}\left(\log \frac{\det(\boldsymbol{\Sigma}')}{\det(\boldsymbol{\Sigma})} + \mathrm{tr}\left(\boldsymbol{\Sigma}'^{-1}\boldsymbol{\Sigma}\right) - b + \left(\boldsymbol{\mu} - \boldsymbol{\mu}'\right)^{\mathrm{T}} \boldsymbol{\Sigma}'^{-1} \left(\boldsymbol{\mu} - \boldsymbol{\mu}'\right)\right). \tag{4.5}$$

There is, however, no such closed form expression for the dissimilarity between two mixtures of Gaussians. Even though various numerical approximations of the KLD between two GMMs have been proposed (Goldberger & Aronowitz, 2005; Hershey & Olsen, 2007), Dautriche (2009; p. 22) converted each GMM $\boldsymbol{g}$ into a single Gaussian distribution $g^\star$ whose mean vector $\boldsymbol{\mu}^\star$ and covariance matrix $\boldsymbol{\Sigma}^\star$ were defined as

$$\boldsymbol{\mu}^\star \equiv \frac{1}{\gamma}\sum_{c=1}^{\gamma} \omega_c \boldsymbol{\mu}_c \qquad \text{and} \qquad \boldsymbol{\Sigma}^\star \equiv \frac{1}{1 - \boldsymbol{\mu}^\star \boldsymbol{\mu}^{\star\mathrm{T}}} \sum_{c=1}^{\gamma} \omega_c \left(\boldsymbol{\Sigma}_c + \boldsymbol{\mu}^\star \boldsymbol{\mu}^{\star\mathrm{T}}\right), \tag{4.6}$$

hence making possible the use of the definition of the KLD between two Gaussians (presented in Equation 4.5) to define the approximation

$$\mathrm{KLD}(\mathbf{g} \,\|\, \mathbf{g}') \approxeq \mathrm{KLD}(g^\star \,\|\, g'^\star) \tag{4.7}$$

where the notational shorthands $\boldsymbol{g}$ and $\boldsymbol{g}'$ stand for $\boldsymbol{g}(\mathbf{x} \mid \Theta)$ and $\boldsymbol{g}(\mathbf{x} \mid \Theta')$, respectively. As the dynamic time warping procedure used by Dautriche is symmetric, using a symmetric extension of the Kullback–Leibler divergence (henceforth SKLD), defined as

$$\mathrm{SKLD}(\mathbf{g}, \mathbf{g}') \equiv \frac{\mathrm{KLD}(\mathbf{g} \,\|\, \mathbf{g}') + \mathrm{KLD}(\mathbf{g}' \,\|\, \mathbf{g})}{2}, \tag{4.8}$$

as the local distance function guarantees the symmetry of the whole acoustic indicator.

Finally, it is worth noting that this defines the acoustic dissimilarity measure we used in Section 3.2 to check the consistency of HTK's output.

10 points go to... Hungary! As previously quoted from the work of Goldberger & Aronowitz (2005), using a mixture model allows for the modeling of arbitrarily shaped densities. However, the very design of Dautriche's single-Gaussian approximation cancels this property. Furthermore, the definitions of approximated parameters $\boldsymbol{\mu}^\star$ and $\boldsymbol{\Sigma}^\star$ are neither discussed in that study nor, to our knowledge, accounted for in the literature. For these reasons, we hereby introduce a novel acoustic indicator of allophony based on a standard and sound approximation of the KLD between two GMMs.

When there is a motivated correspondence between the coindexed components of the GMMs, one can use the following matching-based approximation (Goldberger & Aronowitz, 2005):

$$\mathrm{KLD}(\mathbf{g} \,\|\, \mathbf{g}') \approx \sum_{c=1}^{\gamma} \omega_c \,\mathrm{KLD}(g_c \,\|\, g'_c) \tag{4.9}$$

However, to our knowledge, we can not assume any correspondence between coindexed mixture components in the GMMs produced by HTK (Young et al., 2006; pp. 156–157). Therefore, prior to approximating the KLD between two GMMs as the weighted sum of the KLDs between their components, we need to specify how to pair each component of one GMM with one component of the other. Goldberger & Aronowitz highlighted the fact that any permutation of a given GMM's components yields the exact same mixture of the probability density functions. Put another way, two GMMs $\mathbf{g}$ and $\tilde{\mathbf{g}}$ with equal parameter values but different component ordering actually define the same probability density function. Relying on this property, they defined the following approximation (Goldberger & Aronowitz, 2005; eq. 3):

$$\mathrm{KLD}(\mathbf{g} \,\|\, \mathbf{g}') \approx \min_{\tilde{\mathbf{g}}' \in \wp(\mathbf{g}')} \sum_{c=1}^{\gamma} \varpi_c \,\mathrm{KLD}(g_c \,\|\, \tilde{g}'_c), \tag{4.10}$$

where $\wp(\mathbf{g}')$ denotes the set of all possible permutations of the γ components in the mixture $\mathbf{g}'$, and $\tilde{g}'_c$ denotes the c-th component of the permuted mixture $\tilde{\mathbf{g}}'$. A naive implementation of this matching-based approximation would require searching through all $|\wp(\mathbf{g}')| = \gamma!$ possible permutations of the components (i.e. $17! \approx 3.5 \cdot 10^{14}$ permutations in the case at hand). However, as with any instance of the assignment problem, this search can be solved in only $\mathcal{O}(\gamma^3)$ operations (i.e. $17^3 = 4913$ operations, where $\mathcal{O}$ denotes the asymptotic upper bound on an algorithm's time complexity; Cormen et al., 2001) using the so called Hungarian method (algorithm by Kuhn, 1955, and Munkres, 1957; implementation by Clapper, 2009). Using a symmetrization similar to the one used by Dautriche (2009), we approximate the SKLD between two GMMs $\mathbf{g}$ and $\mathbf{g}'$ as

$$\mathrm{SKLD}(\mathbf{g}, \mathbf{g}') \approxeq \frac{1}{2} \min_{\tilde{\mathbf{g}}' \in \wp(\mathbf{g}')} \sum_{c=1}^{\gamma} \left(\varpi_c \,\mathrm{KLD}(g_c \,\|\, \tilde{g}'_c) + \tilde{\varpi}'_c \,\mathrm{KLD}(\tilde{g}'_c \,\|\, g_c) \right), \tag{4.11}$$

that is to say using the permutation $\tilde{\mathbf{g}}'$ that minimizes the global sum, for each component, of the weighted sums of the component-wise, single-Gaussian KLDs.

Finally, our novel acoustic match-based indicator of allophony, dubbed Ⱥ-MBD, is defined as the sum of the SKLDs between the coindexed emitting states of both phones' HMMs. Formally, its generator function is given by

$$\text{Ⱥ-MBD}(p_i, p_j) \equiv \sum_{\iota \in \{2,3,4\}} \mathrm{SKLD}(\mathbf{g}_{i(\iota)}, \mathbf{g}_{j(\iota)}), \tag{4.12}$$

where $\mathbf{g}_{i(\iota)}$ denotes the GMM associated to the ι-th state of the HMM modeling the phone p_i.

4.2.2 Temporal indicators

Previously defined acoustic indicators rely on the comparison of probability distributions, viz. the GMMs assigned to the emitting states of both phones' HMMs. By doing so, the underlying data they examine are the acoustic features we derived with HTK. However, to a certain extent, temporal information is not immediately accessible in a HMM. Whereas transition probabilities between states constrain the (mostly left-to-right) progression in the observations' structure (i.e. the sequence of acoustic features extracted for each phone), retrieving information about each phone's attested durations in the CSJ from an allophonic HMM is impossible because of the generalization process inherent to any paramater estimation technique.

However, gathering information concerning phone durations might prove to be useful for the discovery of allophony. Broadly speaking, relative durations can indeed be used to define an order on some phoneme classes: the realizations of long vowels tend to be longer than those of short vowels—so we hope—and vowels tend to be longer than consonants (e.g. Grønnum & Basbøll, 2003; Kewley-Port et al., 2007). Thence, comparing two phones' duration distributions may indicate whether or not they are allophones. To our knowledge, duration alone has never been examined a potential cue for allophony. Throughout this study, we test the hypothesis that temporal information can be used to derive effective indicators of allophony.

Defining phonemes in terms of time In this section, we introduce a novel class of indicators of allophony relying on phone duration information, viz. temporal indicators. It is worth noting immediately that, by design, such indicators would fail to account for allophonic processes that affect the duration of their target phonemes (e.g. shortening a long vowel). However, to our knowledge, no such length-affecting process is attested in Japanese.

The main attraction of temporal indicators arises from the simplicity of the data they manipulate—especially compared to the considerable formal apparatus needed for the definition of acoustic indicators. Using the data at hand, information about each sound token's duration is readily available. Indeed, two of our recurrent concerns during the preprocessing operations reported in Section 3.1 concerned temporal annotation. On one hand, we focused on preserving the alignement between the original audio recordings of the CSJ and our custom-built transcription as, in the latter, this alignement is materialized by a sequence of timestamps marking the beginning and the end of each and every phone in the corpus. On the other hand, we successfully checked that all phones in the dataset have non-negative durations.

All temporal indicators to be defined hereafter rely on the same logic at the computational level (Marr, 1982). First, for each phone $p_i \in P$, the exhaustive list of its attested durations are collected in the vector $\mathbf{t}_i$. Then, we assess the temporal dissimilarity between two phones $\{p_i, p_j\} \in P$ by comparing the value of a given statistic about each phone's duration vector $\mathbf{t}_i$ and $\mathbf{t}_j$. More precisely, given a scalar function f, we define the temporal dissimilarity $\mathbb{T}$-f between two phones p_i and p_j as

$$\mathbb{T}\text{-}f(p_i, p_j) \equiv \text{duratio}(p_i, p_j; f) \quad \text{with} \quad \text{duratio}(p_i, p_j; f) \equiv \frac{\min\{f(\mathbf{t}_i), f(\mathbf{t}_j)\}}{\max\{f(\mathbf{t}_i), f(\mathbf{t}_j)\}}, \tag{4.13}$$

where the minimum-over-maximum ratio guarantees the symmetry of all temporal indicators.

We now go on to instantiate four temporal indicators of allophony, specifying a different function f for each of them.

Duration tendency The first two temporal indicators we consider in the present study rely on measures of central tendency, i.e. estimates of the average duration around which a given phone's durations tend to cluster. Though many measures of central tendency have been proposed in the field of descriptive statistics, we only define and evaluate two of them (the inclusion criteria will be discussed hereafter). First, because it is a ubiquitous estimator of central tendency, we define a temporal indicator that relies on a comparison of the arithmetic means of both phones' duration vectors. Referred to as $\mathbb{T}$-mean, its generator function is given by

$$\mathbb{T}\text{-mean}(p_i, p_j) \equiv \text{duratio}(p_i, p_j; \text{mean}) \quad \text{with} \quad \text{mean}(\mathbf{t}) \equiv \frac{1}{|\mathbf{t}|} \sum_{\iota=1}^{|\mathbf{t}|} t_\iota. \tag{4.14}$$

In spite of its popularity, the arithmetic mean has one major limitation: it is highly sensitive to outliers (i.e. extreme values), in the sense that it yields an arbitrarily large estimate of central tendency if the sample contains an arbitrarily large observation. By contrast, the median is one of the most robust estimators of central tendency (Rousseeuw & Croux, 1992, 1993). Therefore, we also introduce a temporal indicator, dubbed $\mathbb{T}$-median, that relies on the comparison of both phones' median duration. Its generator is given by

$$\mathbb{T}\text{-median}(p_i, p_j) \equiv \text{duratio}(p_i, p_j; \text{median}) \quad \text{with} \quad \text{median}(\mathbf{t}) \equiv t_{|t|/2}, \tag{4.15}$$

assuming, for the sake of simplicity, that the values in the duration vector $\mathbf{t}$ are arranged in increasing order.

Duration dispersion The second type of temporal indicators we will examine in the present study rely on measures of statistical dispersion, i.e. estimates of the variability of a given phone's durations. Following the same argument as before, we introduce both a well-known (yet not robust) temporal indicator, and a robust (yet less ubiquitous) indicator. The first dispersion-based temporal indicator we define relies on a comparison of the standard deviations of both phones'

durations. Referred to as $\mathbb{T}$-stdev, its generator function is defined as

$$\mathbb{T}\text{-stdev}(p_i, p_j) \equiv \text{duratio}(p_i, p_j; \text{stdev}) \quad \text{with} \quad \text{stdev}(\mathbf{t}) \equiv \sqrt{\frac{1}{|\mathbf{t}|}\sum_{\iota=1}^{|\mathbf{t}|}(t_\iota - \text{mean}(\mathbf{t}))^2}. \quad (4.16)$$

Because it relies on the arithmetic mean, standard deviation suffers the same sensitivity to outliers. Mirkin (2005; p. 65) recommends using the range (i.e. the difference between the highest and the lowest values) a simple yet effective estimator of dispersion, even though it amounts to comparing the most extreme values in the sample. The generator function of corresponding temporal indicator, dubbed $\mathbb{T}$-range, is straightforwardly given by

$$\mathbb{T}\text{-range}(p_i, p_j) \equiv \text{duratio}(p_i, p_j; \mathbb{T}\text{-range}) \quad \text{with} \quad \text{range}(\mathbf{t}) \equiv \max(\mathbf{t}) - \min(\mathbf{t}). \quad (4.17)$$

Population control As discussed in the first section of this chapter, the major assumption behind indicators of allophony is that their values follow a bimodal distribution whose modes keep apart allophonic pairs from non-allophonic pairs. Consequently, the relative ranks of an indicator's values may matter more than the values themselves. Therefore, any indicator of allophony whose generator function can be defined as a rank-preserving (i.e. monotonic) transformation of another indicator's generator function would yield the exact same observations, at least as far as allophony is concerned. For example, the standard deviation of a given sample is by definition equal to the square root of its variance: as the square root function is monotonic, we know in advance that both statistics would yield identical indicators. For this reason, and for the sake of brevity, we limit our examination of the relevance of temporal information for the discovery of allophony to the aforementioned temporal indicators.

4.2.3 Distributional indicators

Acoustic and temporal indicators of allophony rely on the examination of sound tokens, in the sense that their values depend on the properties of every instance of every allophone of every phoneme in the corpus. In other words, one could argue that our previous assumption of a quantized input (cf. Assumption 3.1) would have been dispensable, as far as acoustic and temporal indicators are concerned. Our answer to this potential objection is that we build upon the work of Peperkamp et al. (2006). As we argued in Section 3.2, only a quantized input allows for the collection of statiscal information about every phone's context distribution, the very data their artificial learner examines.

In this section, we present the various distributional indicators of allophony used throughout the present study, i.e. indicators whose definition is based on a statistical assessment of phone co-occurrence patterns. Even though Peperkamp et al.'s (2006) study contains, chronologically speaking, the first account of a distributional learner of allophony (notwithstanding early work on successor counts by Harris, 1951), we base the following exposition on the extended discussion given in Le Calvez's (2007; pp. 23–52) dissertation.

Complementary distributions The major idea behind Peperkamp et al.'s (2006) model of the acquisition of allophonic rules is to take advantage of the observation that the allophones of a given phoneme have non-overlapping distributions (Peperkamp & Dupoux, 2002). Consider, for example, the aforementioned allophonic palatalization rule in Japanese whereby the alveolar consonant /s/ is realized as an post-alveolar [ʃ] before the high vowel /i/, i.e. s → ʃ / — i (Miller, 1967). By virtue of this rule, the words /miso/ and /mosi/ are to be realized as [miso] ("miso," a traditional Japanese seasoning) and [moʃi] (as in [moʃimoʃi], "hello," when answering a telephone), respectively. Whereas [ʃ] only occurs before (realizations of) /i/, [s] appears in all other contexts. Because, [ʃ] and [s] occur in distinct contexts, they are said to have *complementary distributions* and are thus potential allophones. What is also true for previously discussed linguistic cues of allophony is that complementary distribution is not a sufficient criterion for the discovery of allophony, as highlighted by Peperkamp et al. (2006; p. B33):

> ❝For instance, in French, the semivowel [ɥ] only occurs as the last element in the syllable onset (as *pluie* [plɥi] 'rain'), hence before vowels, whereas the vowel [œ] only occurs in close syllables (as in *peur* [pœʁ] 'fear'), hence before consonants. These two segments, then, have complementary distributions, but in no phonological theory are they considered realizations of a single phoneme.❞

In order to detect phone pairs with complementary distributions (or near-complementary distributions in order to account for noise in the signal or, in the case at hand, smeared distributions in allophonic inventories of high complexity), Peperkamp et al. searched for dissimilar distributions (a strict generalization of complementary distributions, as pointed out by Dunbar, 2009). It is worth emphasizing that the dissimilarity between two context distributions is positively correlated with the likelihood of the two phones being allophones. In other words, highly dissimilar context distributions indicate that the phones are potential allophones.

The first distributional indicator of allophony we use in this study is based on a seminal method of measuring the dissimilarity between two probability distributions, viz. the Jensen–Shannon divergence (henceforth JSD; Lin, 1991). For two arbitrary density functions $f(x)$ and $f'(x)$, their JSD is given by

$$\mathrm{JSD}(f, f') \equiv \frac{\mathrm{KLD}(f \,\|\, f'') + \mathrm{KLD}(f' \,\|\, f'')}{2} \quad \text{with} \quad f''(x) \equiv \frac{f(x) + f'(x)}{2}, \tag{4.18}$$

where KLD is defined as in Equation 4.4. Hence, JSD is defined as a symmetrized measure of the dissimiliraty between each input density function and their average. In the case at hand, probability distributions are discrete distributions and the integral in the definition of KLD can thus be replaced with a summation over all n^2 possible bilateral contexts. Consequently, the generator function of the corresponding distributional indicator (as previously used by Le Calvez, 2007, and Boruta, 2011b), dubbed 𝔻- JSD, is given by

$$\mathbb{D}\text{- JSD}(p_i, p_j) \equiv \frac{1}{2} \sum_{i'} \sum_{j'} f\big(\mathrm{P}(p_i \mid p_{i'}, p_{j'}), \mathrm{P}(p_j \mid p_{i'}, p_{j'})\big) \tag{4.19}$$

where $\mathrm{P}(p_i \mid p_{i'}, p_{j'})$ denotes the probability of the phone p_i occurring after $p_{i'}$ and before $p_{j'}$ (to be rigorously defined hereafter) and the context-wise term is defined as

$$f(x, y) \equiv x \log_2 \frac{2x}{x+y} + y \log_2 \frac{2y}{x+y}. \tag{4.20}$$

Following Cover & Thomas (2006; p. 14), we use the convention that $0 \log 0 = 0$ which is justified by continuity since $x \log x \to 0$ as $x \to 0$. One can prove that JSD's and, consequently, 𝔻- JSD's values lie in $[0, 1]$ (Lin, 1991; p. 148). Various studies have shown that there is no empirical difference between JSD and SKLD for the purpose of learning the allophony relation (Peperkamp et al., 2006; Le Calvez, 2007; Boruta, 2011b). Furthermore, whereas JSD is by definition symmetric and bounded, SKLD is not. For these reasons, 𝔻- JSD is the only information-theoretic distributional indicator that will be used in this study.

The second distributional indicator of allophony we consider relies on a simple, symmetric, and bounded measure of similarity between two probability distributions, viz. the Bhattacharyya coefficient (henceforth BC; Bhattacharyya, 1943). For two arbitrary density functions $f(x)$ and $f'(x)$, their BC is given by

$$\mathrm{BC}(f, f') \equiv \int \sqrt{f(x) f'(x)}\, \mathrm{d}x. \tag{4.21}$$

Following the argument developed for 𝔻- JSD, the generator function of the distributional indicator 𝔻- BC (as used by Le Calvez, 2007, and Boruta, 2011b) is defined as

$$\mathbb{D}\text{- BC}(p_i, p_j) \equiv \sum_{i'} \sum_{j'} \sqrt{\mathrm{P}(p_i \mid p_{i'}, p_{j'})\, \mathrm{P}(p_j \mid p_{i'}, p_{j'})}. \tag{4.22}$$

BC's and, consequently, 𝔻- BC's values lie in $[0, 1]$. If both phones' distributions are strictly identical, then computing their 𝔻- BC amounts to summing over, for each context, the square root of the squared probability of occurrence in this context, that is to say summing to 1. By

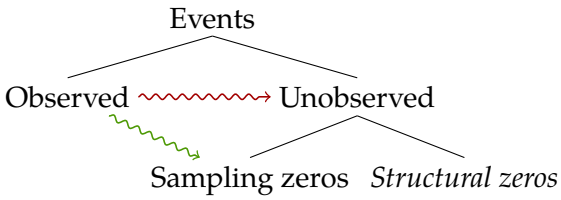


Figure 4.3 — Smoothing: structural zeros vs. sampling zeros. Wavy arrows denote probability mass redistribution during smoothing: whereas LLE and SGT involve redistributing a fraction of the probability mass from the observed events to all unobserved events (here in red), a sound smoothing technique should redistribute probability mass to the sampling zeros only (here in green).

contrast, if each phone is only attested in contexts where the other is not, then each term in the summation is equal to 0, as is the BC of their distributions.

Smoothed distributions As previously mentioned, distributional indicators rely on probability estimates, viz. the probability $\mathrm{P}(p_i \mid p_{i'}, p_{j'})$ of the phone p_i occurring before $p_{i'}$ and after $p_{j'}$. The simplest way to estimate such conditional probabilities from a corpus is to use maximum-likelihood estimates (henceforth MLE). The MLE of the probability $\mathrm{P}(p_i \mid p_{i'}, p_{j'})$ is given by the ratio of the number of occurrences of the phone p_i in this given context (i.e. $p_{i'}$-p_i+$p_{j'}$ using HTK's minus-plus notation, cf. Section 3.2.2), to the number of occurrences of the phone p_i (in any context), that is to say

$$\mathrm{P}(p_i \mid p_{i'}, p_{j'}) \approxeq \frac{\#(p_{i'}\text{-}p_i\text{+}p_{j'})}{\#(p_i)} \tag{4.23}$$

where $\#(\,\cdot\,)$, the count function, denotes the number of occurrences of its argument in the corpus.

If the trigram $p_{i'}$-p_i+$p_{j'}$ is absent from the corpus, the MLE of the probability of the phone occurring in this context is then equal to 0. The problem encountered by Peperkamp et al. (2006; and subsequent studies) is that 𝔻-JSD is only defined if, in every possible context, at least one of the two phones is attested (otherwise the values of the denominators in Equation 4.20 may lead to a division by zero). For this reason, Peperkamp et al.'s computation of the co-occurrence probabilities relied on Laplace–Lidstone estimates (henceforth LLE, a.k.a. add-one smoothing). The LLE of the probability $\mathrm{P}(p_i \mid p_{i'}, p_{j'})$ of the phone p_i occurring before $p_{i'}$ and after $p_{j'}$ is given by

$$\mathrm{P}(p_i \mid p_{i'}, p_{j'}) \approxeq \frac{\#(p_{i'}\text{-}p_i\text{+}p_{j'}) + 1}{\#(p_i) + n^2} \tag{4.24}$$

where n^2 is to be interpreted as the number of possible contexts. Therefore, for all phones and contexts in P, LLEs are stricly positive and 𝔻-JSD can be computed without any complication.

All studies building upon Peperkamp et al.'s experiments have applied the same smoothing technique without further discussion. The contribution of the present research to the study of distributional indicators of allophony consists in assessing how the use of LLE impact these indicators' performance. In order to do so, we compare distributional indicators whose values were computed using LLE to indicators whose values were computed using a technique known as the simple Good–Turing method of frequency estimation (henceforth SGT, algorithm by Gale & Church, 1994, and Gale & Sampson, 1995; implementation by Bane, 2011). Exposing the abundant computational details of SGT is beyond the scope of the present study; suffice it to say that SGT yields frequency-based probability estimates for the various observed events (the attested trigrams in the case at hand), as well as an additional, strictly positive estimate for the total population of unobserved events taken together (all unattested trigrams). SGT does not in itself tell how to share this last quantity between all unseen events (Gale & Sampson, 1995; pp. 218–219); in our experiments, it was uniformly redistributed to each unseen event. For the

sake of completeness, and because it is defined for null probabilities, we will also examine the performance of $\mathbb{D}$- BC under MLE estimates.

It is worth noting, however, that both aforementioned smoothing techniques (viz. LLE and SGT), suffer the same limitation: they reserve some probability mass for *all* unobserved events. Indeed, as discussed by Mohri & Roark (2005; p. 1) and illustrated in Figure 4.3:

> ❝A feature common to all of these techniques is that they do not differentiate between sequences that were unobserved due to the limited size of the sample, which we refer to as *sampling zeros*, from sequences that were unobserved due to their being grammatically forbidden or otherwise illicit, which we call *structural zeros*. Smoothing techniques reserve some probability mass both for sampling and structural zeros.❞

In the case at hand, the consequence is that structural zeros such as phonotactically illicit sequences in the language at hand, e.g. three consonants in a row in Japanese, receive a fraction of the total probability mass. Attempting to identify structural zeros, as investigated by Mohri & Roark, would hinder the comparability of our experiments to those in other studies building upon Peperkamp et al.'s (2006) framework. Therefore, we leave this concern as a recommendation for future research.

Population control As for temporal indicators, additional distributional indicators of allophony were benchmarked during preliminary experiments. These indicators rely on the following measures of (dis)similarity betwen two probability distributions: the symmetrized Kullback–Leibler divergence (Peperkamp et al., 2006; Le Calvez, 2007; Le Calvez et al., 2007; Dautriche, 2009; Martin et al., 2009; Boruta, 2009, 2011b), the Hellinger distance, the Euclidean distance, the Manhattan distance, the Chebyshev distance, and the intersection distance. For the sake of brevity, the performance of these indicators is not reported in the present study.

4.2.4 Lexical indicators

All three aforementioned classes of indicators of allophony (viz. acoustic, temporal, and distributional) are in accordance with Peperkamp et al.'s original, bottom-up hypothesis about the early acquisition of allophony and phonemehood: infants might learn the phonemes of their native language very early in life, before they have a lexicon (Peperkamp & Dupoux, 2002; Peperkamp et al., 2006; Le Calvez et al., 2007).

In this section, we build upon Martin et al.'s (2009)'s subsequent investigation concerning the relevance of lexical (i.e. top-down) information for the acquisition of allophony.

Suddenly, words! Thousands of them! Compared to previously discussed cues of allophony, the typical feature of Martin et al.'s proposal is that it calls for an ancillary word segmentation procedure. Indeed, except for occasional pauses, there are no perceptually salient boundaries between words in fluent—even infant-directed—speech (Kuhl, 2004). Consequently, any model of language acquisition that relies on lexical information has to account for the acquisition of that (preliminary) lexicon, specifying how the continuous stream of speech can be segmented into word-like chunks.

Whereas, as mentioned in Chapter 2, many models of the acquisition of word segmentation have been proposed (e.g. Olivier, 1968; Elman, 1990; Brent & Cartwright, 1996; Christiansen et al., 1998; Brent, 1999; Venkataraman, 2001; Goldwater et al., 2009; Pearl et al., 2010; to cite but a few), Martin et al. approximated infants' emerging lexicon as a set of high-frequency n-grams of phonemes. However, their results suggest that their so called lexical filter, though promising, is less effective when relying on the n-gram approximation than on the reference, dictionary-based word segmentation. In a subsequent study (Boruta, 2011b), we showed that using the lexicon in the output of online models of the acquisition of word segmentation (viz. Brent, 1999; Venkataraman, 2001; Goldwater et al., 2009) also results in a significant deterioration of the performance of a model of the acquisition of allophony based on Peperkamp et al.'s. In

another study (Boruta, 2009; Boruta et al., 2011), we showed that the same models of word segmentation do not resist an increase in allophonic complexity: as the allophonic complexity of their input increases, their performance tends to that of baseline models that insert word boundaries at random in an unsegmented corpus.

For these reasons, we rely on the dictionary-based segmentation provided in the CSJ to compute the values of all lexical indicators to be defined hereafter. Formally, we represent the lexicon $\mathfrak{L}$ as a finite set of words $\mathfrak{L} \equiv \{\mathfrak{w}_1, \ldots, \mathfrak{w}_l\}$, associated to their respective relative frequencies $\mathfrak{F} \equiv \{\mathfrak{f}_1, \ldots, \mathfrak{f}_l\}$. Additionally, let $\mathfrak{L}_\ell$ and $|\mathfrak{w}|$ denote the sublexicon of words of length ℓ and the length of the word $\mathfrak{w}$, respectively.

Terminimal pairs Martin et al.'s lexicon-based learner of allophony is based on the observation that the application of allophonic rules in the vicinity of word boundaries may create minimally different word forms. Consequently, repeating our prior example concerning the allophonic rule of voicing in Mexican Spanish whereby /s/ is realized as [z] before voiced consonants, i.e. [feliz␣nabidad] ∼ [felis␣kumpleaɲos] ("happy Christmas" ∼ "happy birthday"), the intuition is that tracking contrasts on the first or last segment of otherwise identical word forms such as [feliz] ∼ [felis] (which is not a minimal pair stricto sensu) may be relevant in order to learn allophonic rules. In order to distinguish such word pairs from true minimal pairs (i.e. comprising distinct words) such as /felis/ ∼ /relis/ ("happy" ∼ "landslide", also in Mexican Spanish), we will refer to the former as *terminimal pairs*, following the terminology used in a prior study (Boruta, 2009; pp. 38–39).

The more the merrier Martin et al.'s (2009) proposal is actually twofold. Not only do they rely on terminimal pairs, they also rely on word length. As they observed, the frequency distribution of word lengths follows a power law (i.e. most words are very short and few words are very long) and, as a consequence, searching for terminimal pairs in very short words tend to generate many false alarms. In order to account for this observation, Martin et al. set ad hoc bounds on the length of the words they actually considered in their study; for instance, words less than 4 phoneme-long were discarded. More surprisingly, they also introduced an upper bound: words less than 8 phoneme-long were discarded, too, on the grounds that they occur too infrequently to be informative. Although we consider Martin et al.'s argument about the limited informativeness of very shorts words to be acceptable, we do not endorse their views on long words. Whereas it is indeed unlikely that many terminimal pairs be attested on words of, for example, length 10, setting an upper bound on word length amounts to denying oneself the access to precious information in the eventuality of such a lexical contrast. The only word length to be truly irrelevant for the search of terminimal pairs is $\ell = 1$. Indeed, by definition, all words of length $\ell = 1$ form terminimal pairs as they contain no linguistic material apart from their contrasting initial-and-final phone.

The first lexical indicator of allophony to be considered in this study is a mere reformulation of Martin et al.'s so called lexical filter albeit, for the aforementioned reasons, using all words in the lexicon but those comprising a single phone. This indicator, dubbed 𝕃-BTP, relies on a Boolean function whose value for a given phone pair $\{p_i, p_j\} \subseteq P$ is 1 if the lexicon $\mathfrak{L}$ contains at least one terminimal pair of words contrasting on these phones, and 0 otherwise. Formally, its generator function is given by

$$\mathbb{L}\text{-BTP}(p_i, p_j) \equiv [\![\exists\, \{\mathfrak{w}_\iota, \mathfrak{w}_\jmath\} \subseteq \mathfrak{L},\ |\mathfrak{w}_\iota| > 1 \wedge |\mathfrak{w}_\jmath| > 1 \wedge \text{TP}(\mathfrak{w}_\iota, \mathfrak{w}_\jmath, p_i, p_j)]\!] \tag{4.25}$$

where $\text{TP}(\mathfrak{w}, \mathfrak{w}', p, p')$ is 1 if and only if the words $\mathfrak{w}$ and $\mathfrak{w}'$ for a terminimal pair contrasting on the phones p and p'. It is worth noting that this indicator is included in the present study solely for the sake of comparability with Martin et al.'s experiments.

Indeed, we showed (Boruta, 2011b) that the utterly limited number of values that 𝕃-BTP can take, i.e. $\{0, 1\}$, hinders its performance as, for any given phone pair, a single false alarm would switch its value to 1. Therefore, we introduce a simple yet more descriptive lexical indicator of

allophony, dubbed 𝕃-NTP, whose generator function amounts to tallying, for a given phone pair, the number of terminimal pairs contrasting on those phones, that is to say

$$\mathbb{L}\text{-NTP}(p_i, p_j) \equiv \sum_{\ell=2}^{\infty} \sum_{\iota=1}^{|\mathfrak{L}_\ell|} \sum_{\jmath=\iota+1}^{|\mathfrak{L}_\ell|} \text{TP}(\mathfrak{w}_\iota, \mathfrak{w}_\jmath, p_i, p_j). \tag{4.26}$$

Incidentally, 𝕃-BTP's generator function can be defined as a stripped-down version of 𝕃-NTP's by virtue of the following equality:

$$\forall \{p_i, p_j\} \subseteq P,\ \mathbb{L}\text{-BTP}(p_i, p_j) = [\![\mathbb{L}\text{-NTP}(p_i, p_j) > 0]\!]. \tag{4.27}$$

Finally, in order to investigate Martin et al.'s intuition about word length, we also consider a length-weighted lexical indicator of allophony. Referred to as 𝕃-WTP (originally FL*; Boruta, 2011b), its generator function is given by

$$\mathbb{L}\text{-WTP}(p_i, p_j) \equiv \sum_{\ell=2}^{\infty} \ell \sum_{\iota=1}^{|\mathfrak{L}_\ell|} \sum_{\jmath=\iota+1}^{|\mathfrak{L}_\ell|} \text{TP}(\mathfrak{w}_\iota, \mathfrak{w}_\jmath, p_i, p_j). \tag{4.28}$$

Functional load The final indicator of allophony to be considered in this study is based on Hockett's definition of functional load (Hockett, 1955; Surendran & Niyogi, 2006). Functional load is an information-theoretic concept that accounts for the fraction of information content that is lost in a language when a contrast is cancelled out. In the case of allophony, it accounts for the fraction of information content that is lost when the contrast between two given phones is neutralized. As argued by Lyons (1968; p. 82):

> ❝It is to be expected therefore that children will tend to learn first those contrasts which have the highest functional load in the language that they hear about them [...]. The precise quantification of functional load is complicated, if not made absolutely impossible [...].❞

Here, we follow the definitions given by Hockett (1955) and, more recently, Surendran & Niyogi (2006) whereby the information content of the language at hand is represented by its word entropy. Using the available data structures, the language's word entropy $\mathrm{H}(\mathfrak{L})$ is given by

$$\mathrm{H}(\mathfrak{L}) \equiv -\sum_{\iota=1}^{|\mathfrak{L}|} \mathfrak{f}_\iota \log_2 \mathfrak{f}_\iota, \tag{4.29}$$

treating the lexicon as a random variable $\mathfrak{L}$ whose events $\{\mathfrak{w}_1, \ldots, \mathfrak{w}_\mathfrak{l}\}$ have probabilites $\{\mathfrak{f}_1, \ldots, \mathfrak{f}_\mathfrak{l}\}$. The generator function of the corresponding indicator, referred to as 𝕃-HFL (originally HFL; Boruta, 2011b) is then defined as

$$\mathbb{L}\text{-HFL}(p_i, p_j) \equiv \frac{\mathrm{H}(\mathfrak{L}) - \mathrm{H}\big(f(\mathfrak{L}; p_i, p_j)\big)}{H(\mathfrak{L})}, \tag{4.30}$$

where $f(\mathfrak{L}; p_i, p_j)$ denotes Coolen et al.'s (2005) broken-typewriter function. This function returns a new random variable identical to $\mathfrak{L}$ except that p_i and p_j are indistinguishable. $\mathbb{L}\text{-HFL}(p_i, p_j)$ is thus equal to the fraction of information content that is lost when the contrast between p_i and p_j is neutralized. One can prove that 𝕃-HFL's values lie in $[0, 1]$ (Coolen et al., 2005; p. 259).

4.3 Numerical recipes

For the sake of simplicity, and as summarized in Table 4.1, we insignificantly defined indicators of allophony as dissimilarity (𝔸-DTW, 𝔸-MBD, 𝔻-BC-MLE, 𝔻-BC-LLE, and 𝔻-BC-SGT) or similarity (𝕋-mean, 𝕋-median, 𝕋-stdev, 𝕋-range, 𝔻-JSD-LLE, 𝔻-JSD-SGT, 𝕃-BTP, 𝕃-NTP, 𝕃-WTP, and 𝕃-HFL) measures in the preceding section. Moreover, the possible values they can take range from all non-negative reals down to the Boolean set $\{0, 1\}$.

This short section contains an exposition of the various standardization techniques we applied to our indicators of allophony so that a consistent examination of their performance could be developped in the rest of this study.

Table 4.1 — Indicators of allophony.

Indicator	Type	Range	Previous usage
𝔸-DTW	Dissimilarity	$[0, \infty[$	Dautriche (2009)
𝔸-MBD	Dissimilarity	$[0, \infty[$	N/A
𝕋-mean	Similarity	$]0, \infty[$	N/A
𝕋-median	Similarity	$]0, \infty[$	N/A
𝕋-stdev	Similarity	$[0, \infty[$	N/A
𝕋-range	Similarity	$[0, \infty[$	N/A
𝔻-BC-MLE	Dissimilarity	$[0, 1]$	N/A
𝔻-JSD-LLE	Similarity	$[0, 1]$	Le Calvez (2007), Boruta (2011b)
𝔻-JSD-SGT	Similarity	$[0, 1]$	N/A
𝔻-BC-LLE	Dissimilarity	$[0, 1]$	Le Calvez (2007), Boruta (2011b)
𝔻-BC-SGT	Dissimilarity	$[0, 1]$	N/A
𝕃-BTP	Similarity	$\{0, 1\}$	Martin et al. (2009), Boruta (2011b)
𝕃-NTP	Similarity	$[0, \infty[$	N/A
𝕃-WTP	Similarity	$[0, \infty[$	Boruta (2011b)
𝕃-HFL	Similarity	$[0, 1]$	Boruta (2011b)

4.3.1 Turning similarities into dissimilarities

The first transformation we applied to all aforementioned indicators of allophony consisted in turning similarities into dissimilarities, so that the likelihood of two phones $\{p_i, p_j\} \subseteq P$ being allophones is always inversely proportional to the value δ_{ij} any indicator maps them to.

To this aim, we merely flipped all similarity-based indicators, substracting each indicator's individual value to the indicator's maximum value (Vakharia & Wemmerlöv, 1995; Esposito et al., 2000). Let δ_{ij} denote the original similarity rating and δ'_{ij} denote the dissimilarity between the phones p_i and p_j in the flipped dissimilarity matrix $\mathbf{\Delta}'$, we have

$$\delta'_{ij} \equiv \left(\max_{i' < j'} \delta_{i'j'} \right) - \delta_{ij}. \tag{4.31}$$

This minimalistic transformation preserves the range of the values, the relative dis/similarities, as well as the shape of the values' distribution.

4.3.2 Standardizing indicators

Because they rely on the aforestated compactness hypothesis, the behavior of many classification and clustering algorithms depends on input dissimilarity ratings, proximity scores, or the minimum distance rule (Mirkin, 2005; pp. 64–70). Such quantities, e.g. the pervasive Euclidean distance, are sensitive to differences in the scale of values: features with large values will have a larger influence than those with small values. Equal contribution of all features to the classification or the clustering may or may not be a desirable property but, as this numerical influence does not necessarily reflect their real importance for the discovery of classes, and in the absence of a priori knowledge of the relative weight of each feature, we follow Mirkin's (2005) argument that standardization (a.k.a. normalization) of the values is a reasonable, if not necessary, preprocessing step.

The most common way to standardize data is to apply a linear transformation (so that relative proximities remain intact) to the raw data, by first shifting the values by a measure of central tendency, and then rescaling them by a measure of statistical dispersion. For instance, the so called z-score standardization has been very popular in data mining and computational linguistics (including in models of the acquisition of allophony; e.g. Peperkamp et al., 2006).

It relies on shifting the values by their arithmetic mean and rescaling them by their standard deviation; yet, as argued by Mirkin (2005; p. 65):

> ❝Thus standardized, contributions of all features to data scatter become equal to each other because of the proportionality of contributions and standard deviations. [However], by standardizing with standard deviations, we deliberately bias data in favor of unimodal distributions, although obviously it is the bimodal distribution that should contribute to clustering most.❞

This observation is especially true in the case at hand, where features are the aforementioned indicators of allophony. Indeed, as stated in Section 4.1.2, our goal is to discover bimodal distributions in the input data, distributions whose modes emblematize the dichotomy between allophonic and non-allophonic pairs of phones.

Range standardization In order to circumvent the limitations inherent to the use of standard deviation as the scaling factor, Mirkin (2005) recommends using the values' range as an appropriate scaling factor. Hence, the first standardization technique we consider in the present study consists in first shifting by the minimum value and then rescaling by the range. Let δ'_{ij} denote the dissimilarity between the phones p_i and p_j in the range-standardized dissimilarity matrix $\mathbf{\Delta}'$, we have

$$\delta'_{ij} \equiv \frac{\delta_{ij} - (\min_{i'<j'} \delta_{i'j'})}{(\max_{i'<j'} \delta_{i'j'}) - (\min_{i'<j'} \delta_{i'j'})}. \tag{4.32}$$

Using this standardization technique, both relative dissimilarities and the shape of the distributions are preserved, but all values lie in the same range, i.e. $[0, 1]$.

Ranks standardization Because of the assumptions underlying Peperkamp et al.'s (2006) framework, we introduce another standardization technique that we will refer to as rank standardization. All things considered, the main assumption behind what we refer to as indicators of allophony is indeed that allophonic pairs should be more similar than non-allophonic pairs. Put another way, allophonic pairs should rank lower than non-allophonic pairs in the arrangement of all phone pairs $\{p_i, p_j\} \subseteq P$ by increasing order of the values to which a given indicator maps them. Thus, a transversal problem (that we address is this chapter and the next) is to determine whether allophony and phonemehood can be learned by a mere ranking of the candidate pairs, or if the actual dissimilarity ratings denoted by each indicator's values are meaningful. To this aim, we will compare the results obtained with both range- and ranks-based standardizations. Let δ'_{ij} denote the dissimilarity between the phones p_i and p_j in the ranks-standardized dissimilarity matrix $\mathbf{\Delta}'$. In the general case, we have

$$\delta'_{ij} \equiv 1 + \left|\left\{\{p_{i'}, p_{j'}\} \subseteq P : \delta_{i'j'} < \delta_{ij}\right\}\right|, \tag{4.33}$$

that is to say the rank of the pair $\{p_i, p_j\}$ is given by 1 plus the number of pairs with stricly lesser values. In the case of tied values, the ranks of the corresponding pairs are averaged (Jones et al., 2001; cf. `scipy.stats.mstats.rankdata`); for example, if the minimal value a given indicator can take is attested for two distinct phone pairs, both pairs would receive the pseudorank of $(1 + 2)/2 = 1.5$.

4.3.3 Addressing the frequency effect

As highlighted in Sections 3.1.3 and 3.2, various frequency effects can be observed in our data: on one hand, the distribution of phoneme frequencies follows a power law (cf. Figure 3.2) and, on the other hand, the number of allophones HTK computes for a given phoneme appears to be, at any allophonic complexity, positively correlated with that phoneme's frequency (cf. Table 3.7).

However, having equal expectations for the performance of a computational model on the least and the most frequent instances of the problem would be unrealistic. Indeed, training any such model on the CSJ implies examining 50 times less evidence for the allophones of /iː/

than for the allophones of /a/. Notwithstanding potential sampling issues in the constitution of the CSJ, we consider this lack of balance in our data to be an inherent property of the Japanese language; yet, we will account for this imbalance by assigning a frequency-based weight to each phone pair $\{p_i, p_j\}$ in the allophonic inventory P.

Individual weights The weighting scheme we introduce in the present study relies on mere frequency counts. Concretely, for a given allophonic inventory P, the weight $w_i \in \mathbb{N}$ we assign to each phone $p_i \in P$ is given the frequency of occurrence of p_i in the corpus, i.e. $w_i \equiv \#(p_i)$ (cf. Section 4.2.3). As all allophones were computed by HTK on the basis of the actual content of the CSJ (rather than using an a priori allophonic grammar of Japanese), all individual weights are, by definition, strictly positive. As a matter of convenience, let $\mathbf{w}$ denote the vector of length n collecting all individual weights, i.e. $\mathbf{w} \equiv (w_1, \ldots, w_n)$.

Because such weights are estimated by the phones' occurrence frequencies, they can be interpreted as instance weights. Repeating our prior analogy with canopy clustering (McCallum et al., 2000), any allophonic inventory can be considered as the output of a canopy clustering technique whereby the inherent variability in human speech was reduced to a finite, quantized set of objects (i.e. the allophones) complemented with information about the fraction of variability accounted for by each canopy (i.e. the phones and their respective weights, the latter indicating how many sound tokens were used to estimate the parameters of each allophonic HMM).

Compared to previous studies building upon Peperkamp et al.'s (2006) and, especially, those where a systematic examination of the influence of the input's allophonic complexity was conducted (Martin et al., 2009; Boruta, 2009, 2011b), using frequency-based instance weights will facilitate the comparability between a given model's outputs at various complexities. Indeed, another consequence of using frequency-based instance weights is that the total number of instances (i.e. $\sum_i w_i$) is constant across all possible allophonic complexities, as it is equal to the overall number of sound tokens in the input corpus.

Pairwise weights In Peperkamp et al.'s framework, the core object to be manipulated is a pair of phones. Therefore, we propose the following straightforward extension of our frequency-based definition of weights to phone pairs. We define the weight $w_{ij} \in \mathbb{N}$ assigned to the phone pair $\{p_i, p_j\} \subseteq P$ as the product of both phones' individual weights, i.e. $w_{ij} \equiv w_i w_j$. As a matter of convenience, let $\mathbf{W}$ denote the symmetric $n \times n$ weight matrix $\mathbf{W} \equiv [w_{ij}]$ in which all pairwise weights are collected.

Because the individual weight w_i is given by the number of sound tokens that were canopy-clustered as the phone p_i, the pairwise weight $w_{ij} \equiv w_i w_j$ can also be interpreted as an instance weight: indeed, w_{ij} is equal to the number of sound pairs in which one sound token was canopy-clustered as p_i and the other as p_j. For each tested allophonic complexity, the distribution of the pairwise weights is illustrated in Figure 4.4, using a logarithmic scale to account for the great variability between weights computed for pairs consisting in two high-frequency phones and for pairs consisting in two low-frequency phones. The box plots in Figure 4.4 suggest that, whereas the median pairwise weight decreases as the allophonic complexity increases, pairwise weights' attested range enlarges until it plateaus, from $^{782}/_{25}$ allophone per phoneme onwards.

There's no such thing as a free lunch As far as allophony is concerned, the introduction of pairwise weights in the framework has one major consequence: it alters the ratio of the number of allophonic pairs to the number of non-allophonic pairs. In order to quantify the influence of pairwise weights on class balance, we computed two such ratios. On one hand, we computed the allophonic to non-allophonic ratio as would be observed by Peperkamp et al.'s (2006) learner, i.e.

$$\text{class-balance}(\mathbf{A}^*) \equiv \frac{\sum_{i<j} [\![a^*_{ij} = 1]\!]}{\sum_{i<j} [\![a^*_{ij} = 0]\!]}, \tag{4.34}$$

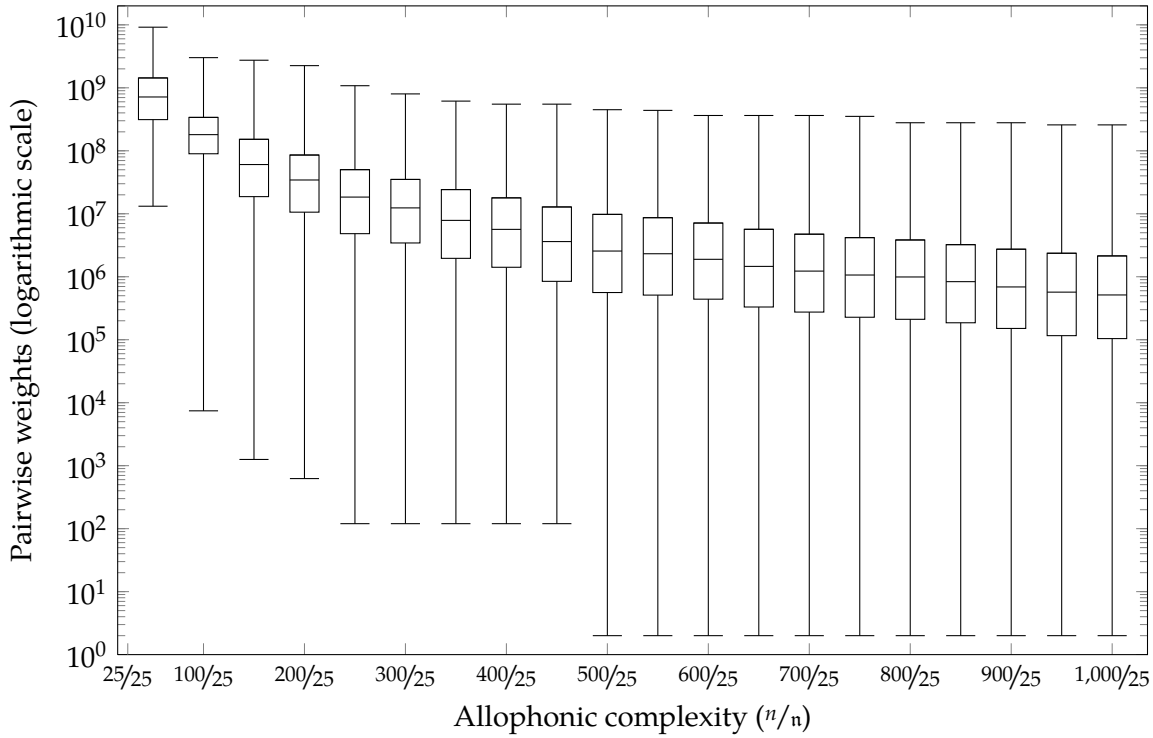


Figure 4.4 — Box plots of the computed pairwise weights for each allophonic complexity. Each box plot depicts the weights **W** observed at the corresponding allophonic complexity, using the following descriptive statistics: minimum value (the lower whisker), maximum value (the upper whisker), first quartile (the bottom of the box), upper quartile (the top of the box), and median value (the band inside the box).

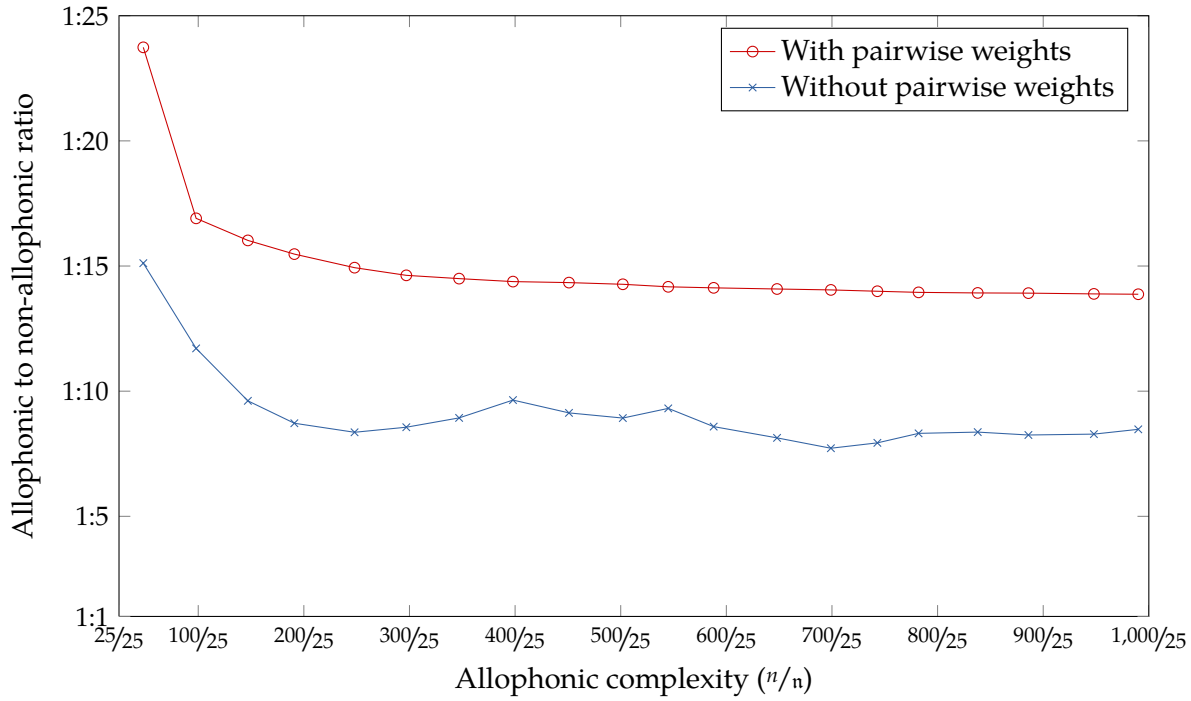


Figure 4.5 — Influence of the introduction of pairwise weights on class balance, given by the ratio of allophonic pairs to non-allophonic pairs as a function of allophonic complexity.

and, on the other hand, we computed the updated allophonic to non-allophonic ratio, as will be observed by our learner—that is to say, when pairwise weights are available, i.e.

$$\text{class-balance}(\mathbf{A}^*, \mathbf{W}) \equiv \frac{\sum_{i<j} w_{ij} [\![a^*_{ij} = 1]\!]}{\sum_{i<j} w_{ij} [\![a^*_{ij} = 0]\!]}. \tag{4.35}$$

The results are presented in Figure 4.5 as a function of the allophonic complexity of the tested inventories. Notwithstanding the lowest allophonic complexities such as $^{48}/_{25}$ or $^{98}/_{25}$ allophones per phoneme, these ratios appear to be quite unchanging: while, broady speaking, not using frequency information yields a ratio of the number of allophonic pairs to the number of non-allophonic pairs approximately equal to 1:8, using frequency-based pairwise weights yields a ratio approximately equal to 1:14.

As will be further discussed in Chapter 5, one limitation of Peperkamp et al.'s framework is that the learner has to process all possible phone pairs to discover which ones are allophones. In the case at hand, non-allophonic pairs are therefore useless, as no relation brings such phones together. Therefore, whereas an artificial learner à la Peperkamp et al. has to reject, on average, 7 out of every 8 phone pairs, an artificial learner provided with frequency information would have to reject 13 out of every 14 phone pairs. Hence, accounting for the power law in phoneme frequencies comes at the cost of increasing the proportion of negative objects in the input data, hence the title of this paragraph.

4.4 Prognoses of allophony

This section contains a preliminary examination of indicators' performance for the discrimination of allophonic pairs vs. non-allophonic pairs. This examination is actually twofold. First, we replicate part of Martin et al.'s (2009) experiments using a probabilistic assessment of class separation that quantifies, for a given indicator of allophony, the probability of non-allophonic pairs being more dissimilar than allophonic pairs. Second, we introduce a novel visualization technique allowing for the exploration of the classwise distribution of an indicator's values.

Our focus here is on emphasizing the distinction between such tools, that we refer to as prognoses, and true evaluation methods. Whereas the purpose of an evaluation method is to judge the output of a given model or algorithm on a given task according to given quality criteria, the purpose of a prognosis method is to inspect the data in order to provide a better understanding of its inner content and structure, and to hint at transformations or algorithms that would be relevant for the task at hand. In other words, a prognosis method does not suggest conclusions: it supports decision making and, at best, it suggests expectations.

4.4.1 Rank-sum test of class separation

The first prognosis of allophony we present is a rank-sum test, originally proposed by Herrnstein et al. (1976), and first used in the context of allophony learning by Martin et al. (2009). Given two observation samples, this non-parametric test aims at assessing whether one of the two samples tends to have larger values than the other. In the case at hand, we test the hypothesis that non-allophonic pairs tend to be more dissimilar than allophonic pairs.

Definitions Herrnstein et al.'s (1976) statistic, denoted ρ, is equal to the probability of non-allophonic pairs being more dissimilar than allophonic pairs. Let $\mathbf{A}^*$ and $\mathbf{\Delta}$ denote, respectively, the reference allophony matrix and the dissimilarity matrix collecting the values of a given indicator of allophony. Concretely, ρ is a MLE of the probability that a randomly-drawn non-allophonic pair will score higher than a randomly-drawn allophonic pair in $\mathbf{\Delta}$. We first give a definition of ρ without taking instance weights into account, i.e. as used by Martin et al. (2009).

In this case, we have

$$\rho(\boldsymbol{\Delta}; \mathbf{A}^*) \equiv \frac{\sum_{i<j} \sum_{i'<j'} [\![a^*_{ij} = 0]\!][\![a^*_{i'j'} = 1]\!][\![\delta_{ij} > \delta_{i'j'}]\!]}{\left(\sum_{i<j} [\![a^*_{ij} = 0]\!]\right)\left(\sum_{i'<j'} [\![a^*_{i'j'} = 1]\!]\right)}. \tag{4.36}$$

To our knowledge, no formulation of ρ has been proposed so far to account for weighted observation samples. Using the frequency-based instance weights collected in $\mathbf{W}$, as defined in the previous section, we propose the following weighted extension of the definition of ρ:

$$\rho(\boldsymbol{\Delta}; \mathbf{A}^*, \mathbf{W}) \equiv \frac{\sum_{i<j} \sum_{i'<j'} w_{ij} w_{i'j'} [\![a^*_{ij} = 0]\!][\![a^*_{i'j'} = 1]\!][\![\delta_{ij} > \delta_{i'j'}]\!]}{\left(\sum_{i<j} w_{ij} [\![a^*_{ij} = 0]\!]\right)\left(\sum_{i'<j'} w_{i'j'} [\![a^*_{i'j'} = 1]\!]\right)}. \tag{4.37}$$

This amended definition is consistent with our conception of weights as instance weights: w_i was defined as the number of sound tokens that were canopy-clustered as p_i, $w_{ij} \equiv w_i w_j$ was defined as the number of pairs of sound tokens in which one sound was canopy-clustered as p_i and the other as p_j, and the product $w_{ij} w_{i'j'}$ in the numerator of Equation 4.37 is equal to the number of pairs of pairs of sound tokens in which the first pair was canopy-clustered as $\{p_i, p_j\}$ and the other as $\{p_{i'}, p_{j'}\}$.

It is worth mentioning that these definitions of ρ with a double summation are given here for the sake of simplicity: a straightforward implementation of these definitions would however require $\mathcal{O}(n^4)$ operations, as it would consist in iterating over all possible pairs of pairs of phones; yet, because Herrnstein et al.'s test relies on ranks, the time complexity of the computation of ρ can be reduced to $\mathcal{O}(n^2 \log_2 n)$ by first copying all pairwise dissimilarity ratings (the lower triangle in $\boldsymbol{\Delta}$) into a list, sorting this list—hence the linearithmic complexity—and then working on the ranks of the values in the sorted list (Jones et al., 2001; cf. `scipy.stats.mstats.mannwhitneyu`).

Being probability estimates, all ρ values (weighted or not) lie in $[0, 1]$ and $\rho = .5$ indicates chance. Whereas high and low values both indicate a significant separation of the two allophonic statuses in $\boldsymbol{\Delta}$, it is worth noting that values strictly below .5 indicate a reverse separation of statuses; in that case, allophonic pairs tend to be mapped to larger dissimilarity ratings than non-allophonic pairs and, thus, the indicator is in fact a similarity measure. While indicators with $\rho < .5$ could be turned into proper indicators of allophony by flipping their values (cf. Section 4.3.1), indicators with $\rho = .5$ simply contain no relevant information for the purpose of separating allophonic pairs from non-allophonic ones. For instance, a uniform, random drawing of dissimilarity ratings would yield a ρ value equal to .5.

Chances of allophony Herrnstein et al.'s rank-sum test is related to other statistics, viz. the area under the receiver operating characteristic curve (a.k.a. the AUC) and the Mann–Whitney U test. All these statistics, however, should not be considered as appropriate evaluation measures in the case at hand, as they do not rely on a partition of all possible phone pairs into allophonic and non-allophonic pairs. Indeed, although we ambiguously argued in a previous study that ρ may be used to evaluate the performance of an indicator across all possible discrimination thresholds (Boruta, 2011b; p. 2), its limitation lies in the fact that no explicit discrimination threshold is actually applied or even searched for. For this reason, ρ is nothing but a prognosis of allophony: it is a probabilistic assessment of how effective a given indicator may be, not a complete algorithm that decides which pairs of phones are allophones and which are not.

Results Weighted ρ values are presented in Figures 4.6, 4.7, 4.8, and 4.9 as a function of allophonic complexity for acoustic, temporal, distributional, and lexical indicators, respectively. Let us mention immediately that, throughout the present study, the y-axes of all plots illustrating prognoses or evaluation measures are annotated with an arrow indicating how the particular measure should be interpreted: ↑ for measures that are positively correlated with the quality of the underlying data (i.e. the higher the better), and ↓ for measures that are inversely correlated with the quality of the underlying data (the lower the better).

As presented in Figure 4.6, the prognosis of class separation for acoustic indicators is encouraging. First and foremost, both acoustic indicators' prognosis of allophony is significantly above

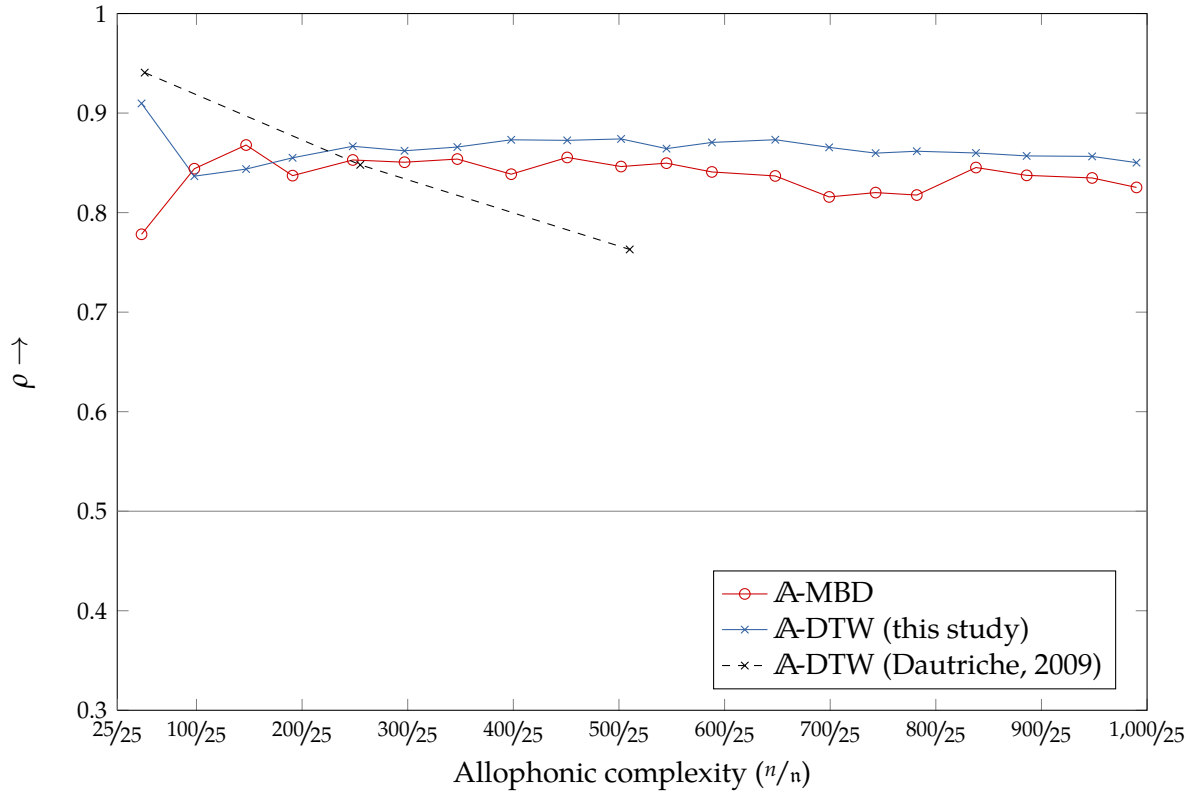


Figure 4.6 — Prognosis of class separation by the ρ statistic for acoustic indicators of allophony, as a function of allophonic complexity. The gray line indicates chance.

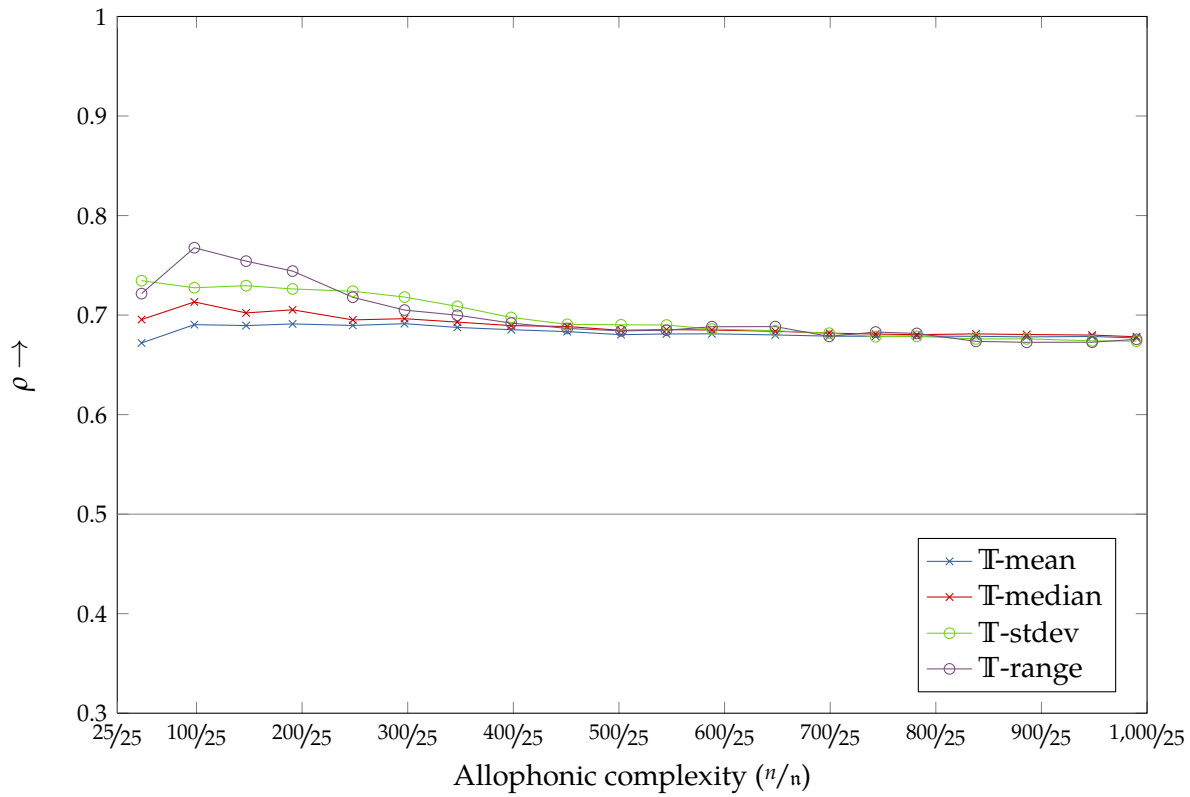


Figure 4.7 — Prognosis of class separation by the ρ statistic for temporal indicators of allophony, as a function of allophonic complexity. The gray line indicates chance.

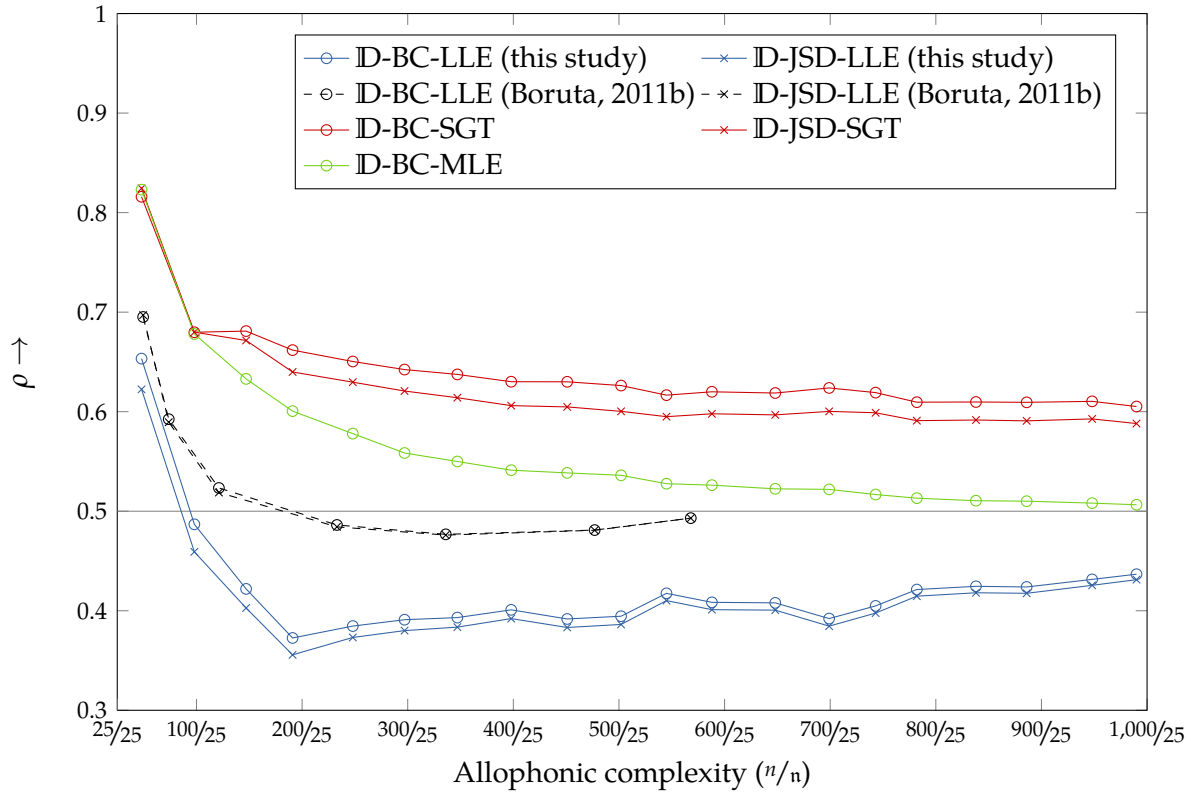


Figure 4.8 — Prognosis of class separation by the ρ statistic for distributional indicators of allophony, as a function of allophonic complexity. The gray line indicates chance.

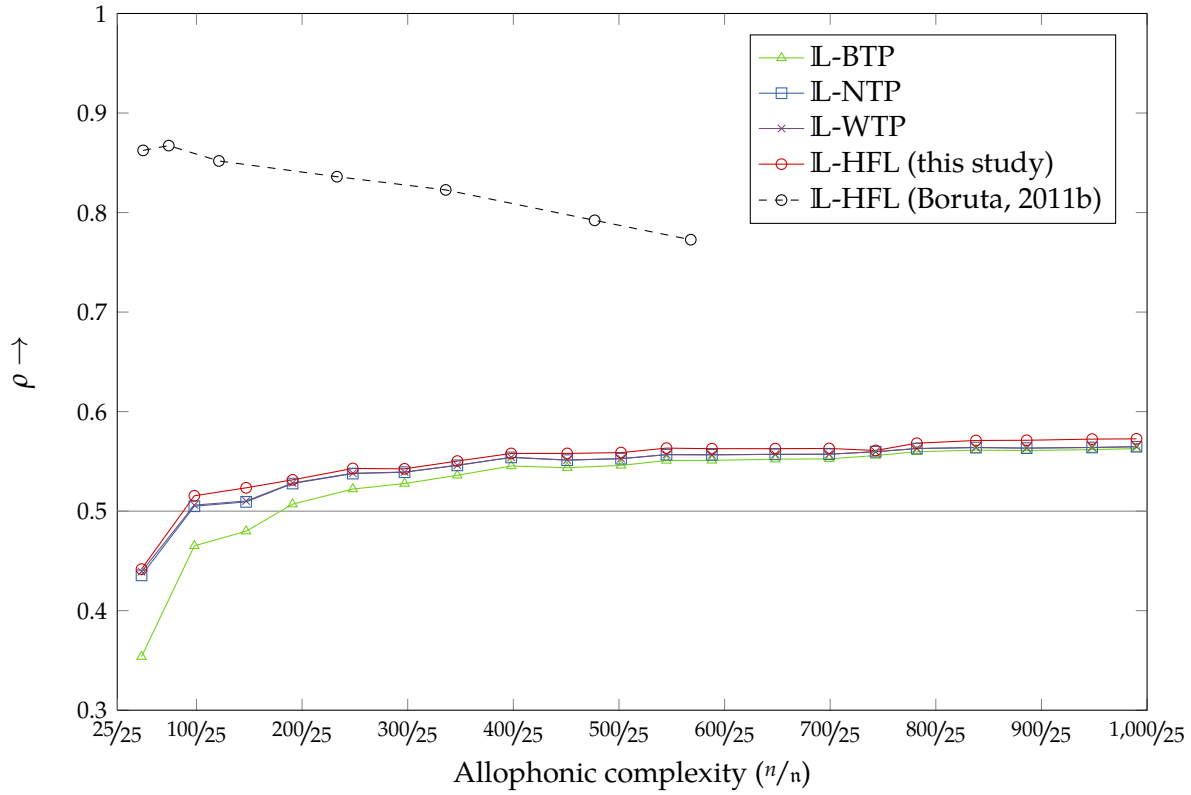


Figure 4.9 — Prognosis of class separation by the ρ statistic for lexical indicators of allophony, as a function of allophonic complexity. The gray line indicates chance.

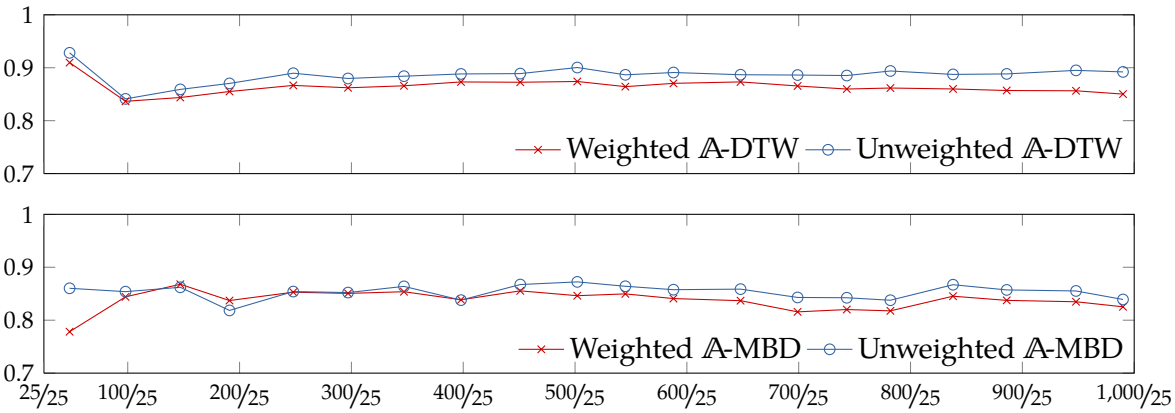


Figure 4.10 — Unweighted vs. weighted ρ prognosis for acoustic indicators.

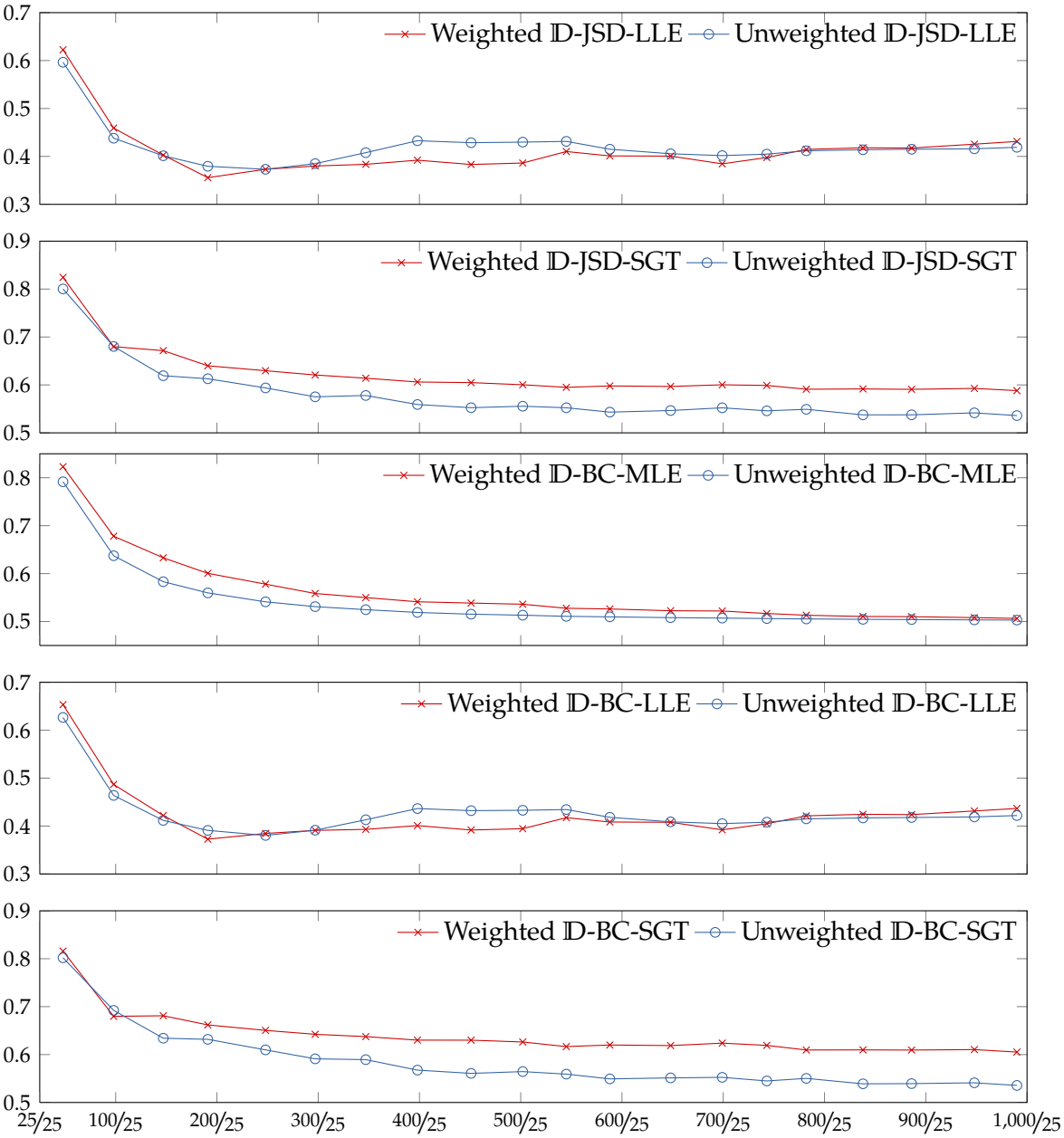


Figure 4.11 — Unweighted vs. weighted ρ prognosis for distributional indicators.

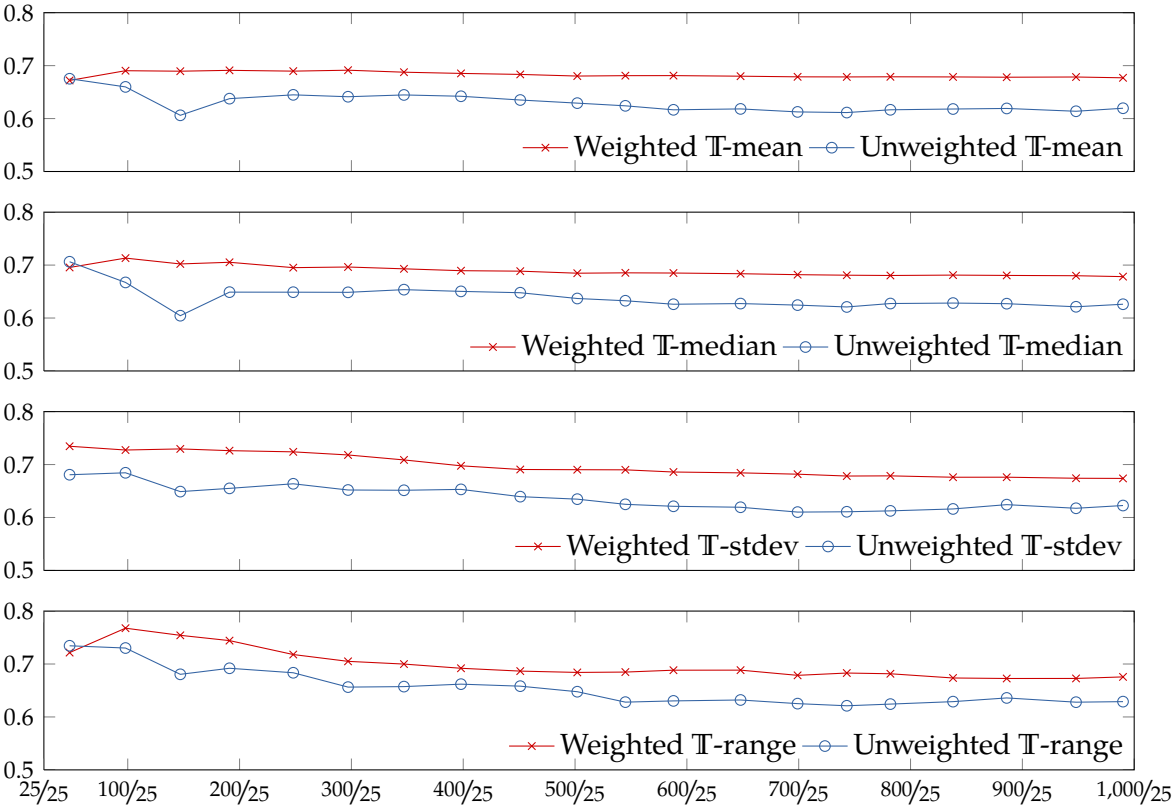


Figure 4.12 — Unweighted vs. weighted ρ prognosis for temporal indicators.

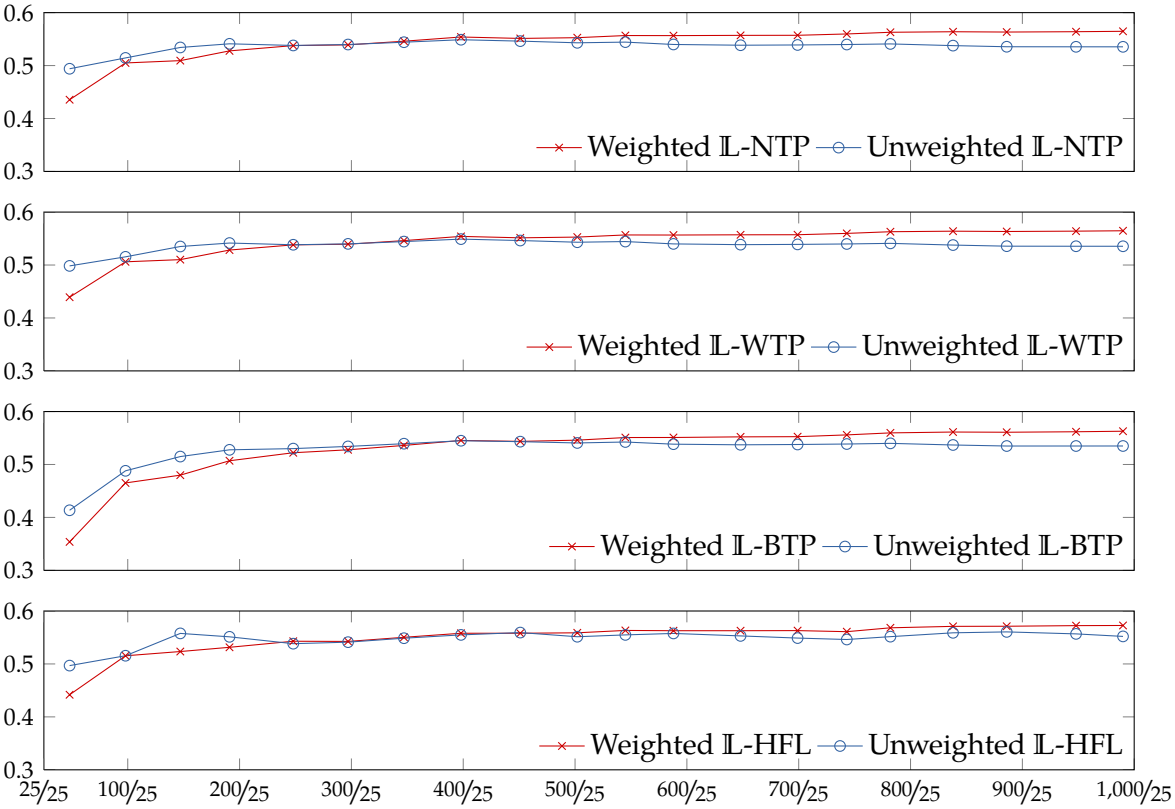


Figure 4.13 — Unweighted vs. weighted ρ prognosis for lexical indicators.

chance. Moreover, 𝔸-DTW and 𝔸-MBD seem to be resistant (if not insensitive) to an increase in allophonic complexity. Indeed, as the complexity of the input increases, ρ values tend to lie in the $[.8, .9]$ band, indicating that there is more than 8 chances in 10 of a randomly-drawn non-allophonic pair being acoustically more dissimilar than a randomly-drawn allophonic pair. Therefore, despite the fact that allophones were computed by HTK on the basis of the acoustic dissimilarity between their realizations, it appears that any given phoneme's allophones are still, on average, acoustically homogeneous. Because they were derived from a comparable corpus—viz. the CSJ, albeit using unmatchable preprocessing choices—we also report in Figure 4.6 the ρ values obtained by Dautriche (2009) for 𝔸-DTW. These three points suffice to suggest that, in that setup, 𝔸-DTW does not resist an increase in allophonic complexity. Because both studies rely on the identical definitions of 𝔸-DTW, this discrepancy is necessarily due to Dautriche's transcription scheme. Indeed, in that study, allophonic HMMs were trained for an inventory of 49 phonemic categories in which, for example, allophonic palatalizations were kept transcribed as-is. We suspect that, because of this subphonemic transcription scheme, Dautriche's experiments ran into the data sparseness problem we circumvented in Section 3.1: all other things being equal, doubling the number of phoneme-level categories can only scatter the information available in the training corpus.

Although less impressive than that of the acoustic indicators, the prognosis of class separation for temporal indicators is surprisingly good, especially considering the limited information on which these indicators rely. As presented in Figure 4.7, all four indicators' ρ values tend to $\rho \approx .7$ as the allophonic complexity of their input increases. Indeed, 𝕋-mean, 𝕋-median, 𝕋-stdev, and 𝕋-range hardly suffer an increase in allophonic complexity and are virtually indistinguishable from $^{450}/_{25}$ allophones per phoneme onwards. Not only do these indicators of allophony share the same underlying logic, they also appear to share the same behavior, regardless of the exact definition of their ancillary statistic and, more surprisingly, no matter whether the focus is on comparing the central tendency or the statistical dispersion of both phones' durations. It is worth noting, however, that indicators relying on measures of statistical dispersion yield a slightly better separation of allophonic statuses at lower allophonic complexities.

While the behavior of both acoustic and temporal indicators of allophony looks promising, the same cannot be said for distributional indicators, as presented in Figure 4.8. Indeed, all five indicators' prognoses drop significantly for the first increases in allophonic complexity. Martin et al. (2009) argued that this is due to the fact that, in corpora of high allophonic complexity, every phone has an extremely narrow distribution and, hence, having complementary (or, to say the least, dissimilar) distributions is the rule rather than the exception. Notwithstanding this sensitivity to allophonic complexity, the major observation is that while the very definition of the measure used to compare both phones' context distributions (viz. JSD or BC) appear to have virtually no effect on the prognosis of class separation, the technique used to estimate the probabilities has a significant impact. Indeed, three groups emerge in Figure 4.8: the indicators relying on SGT estimates whose ρ tend to values slightly above .6, the indicator relying on MLEs whose ρ tend to chance, and the indicators relying on LLEs whose ρ drops below chance from $^{98}/_{25}$ allophones per phoneme onwards. This last observation corroborates previously reported ρ values for distributional indicators relying on LLEs (Martin et al., 2009; Boruta, 2011b); yet, whereas we previously observed (Boruta, 2011b) that the prognosis of class separation for indicators relying on LLEs of SKLD, JSD, or BC indeed drops below chance as the complexity of their input increases (as reported in Figure 4.8 for comparable indicators), Martin et al.'s (2009) figures suggest that their SKLD-based indicator plateaus slightly above chance. We are unfortunately unable to account for this discrepancy except, maybe, by an effect of the very different processes used to create the allophonic inventories (linguistic rules vs. random partitions, as presented in Section 3.2.1). Furthermore, while assessing the numerical discrepancies between LLE- and SGT-based smoothing would be intractable—as it amounts to the examination of thousands of probability distributions—this shallow comparison is sufficient to show that Peperkamp et al.'s (2006) use of smoothing as a numerical recipe to avoid null

probabilities has a significant impact on the separation of the classes. Indeed, focusing on the behavior of 𝔻-BC-MLE, 𝔻-BC-LLE, and 𝔻-BC-SGT, Figure 4.8 suggests that whereas SGT and LLE yield contradictory observations as the complexity increases, ρ values for MLE (whereby no structural zero may receive some probability mass) suggest that distributional indicators simply tend to contain no relevant information to the discovery of allophony.

Finally, the graphs presented in Figure 4.9 suggest that all four lexical indicators contain little information about the dichotomy between allophonic and non-allophonic pairs. Whereas, as we observed for temporal indicators, the specific formulation of each indicator's generator function appears to have virtually no influence on the separation of the classes, a remarkable property of lexical indicators is that, contrary to other classes, their ρ values tend to increase with the allophonic complexity of their input. Moreover, their prognosis of allophony is below chance for the lowest complexities, i.e. up to $^{191}/_{25}$ allophones per phoneme, meaning that the lexical alternations they keep track of are not due to allophonic processes, but to true minimal pairs. However, whereas our results confirm that the coarse-grained definition of Martin et al.'s 𝕃-BTP impedes its discrimination power, such generally low values are inconsistent with those reported in previous studies (Martin et al., 2009; Boruta, 2011b; comparable results from the latter study are reported in Figure 4.9, i.e. the prognosis of allophony for 𝕃-HFL relying on the orthographic segmentation). We suspect that this is due to the unlikeness of the underlying lexicons: whereas both aforementioned studies used lexicons derived from infant-directed corpora, we rely on the miscellaneous lexicon of an adult-directed corpus that comprises, for example, academic presentations and public speeches.

For the sake of completeness, Figures 4.10, 4.11, 4.12, and 4.13 illustrate the influence of using instance weights on the prognosis of allophony for acoustic, distributional, temporal, and lexical indicators, respectively. Whereas no definite pattern emerges from these results—except, perhaps, for temporal and SGT-based distributional indicators which appear to slightly benefit from the introduction of weights—the major observation is that there is no significant difference in the ρ values computed for previous (unweighted) and the present (weighted) approaches to the discovery of allophony. In each figure, both curves follow the same trend as the allophonic complexity of the input increases. Such an observation is particularly interesting when compared with the class-balance ratios presented in Figure 4.5: we indeed observed that introducing frequency-based instance weights in Peperkamp et al.'s framework almost doubles the ratio of non-allophonic to allophonic pairs the artificial learner has to consider; yet, for all indicators, it appears to have no significant effect on the probability of non-allophonic pairs being more dissimilar than allophonic pairs.

For this reason, and because they allow us to account for the power law governing phonemes' frequency distribution, frequency-based instance weights will be used in all experiments reported from this point onwards.

4.4.2 Combining indicators of allophony

Up to this point, indicators of allophony were considered in isolation. However, as discussed in Section 4.2, not a single one of the four types of indicator we study in the present research is a sufficient cue for the discovery of allophony. Indeed, not all acoustically, temporally, distributionally, or lexically similar phones are allophones. Consequently, most experiments carried out under Peperkamp et al.'s framework combined various indicators of allophony: distributional and articulatory (Peperkamp et al., 2006; Le Calvez, 2007; Le Calvez et al., 2007), distributional and lexical (Martin et al., 2009; Boruta, 2009, 2011b), distributional, acoustic, and articulatory (Dautriche, 2009), and distributional, acoustic, temporal, and lexical (this study; cf. Table 3.4). However, the issue of supplementing Peperkamp et al.'s framework with a general and flexible combination mechanism has seldom been discussed. In a previous study, we specifically addressed the issue of the combination of indicators, but failed to propose an effective reformulation of Peperkamp et al.'s learner—none of the combination schemes we tested (conjunctive vs. disjunctive, and

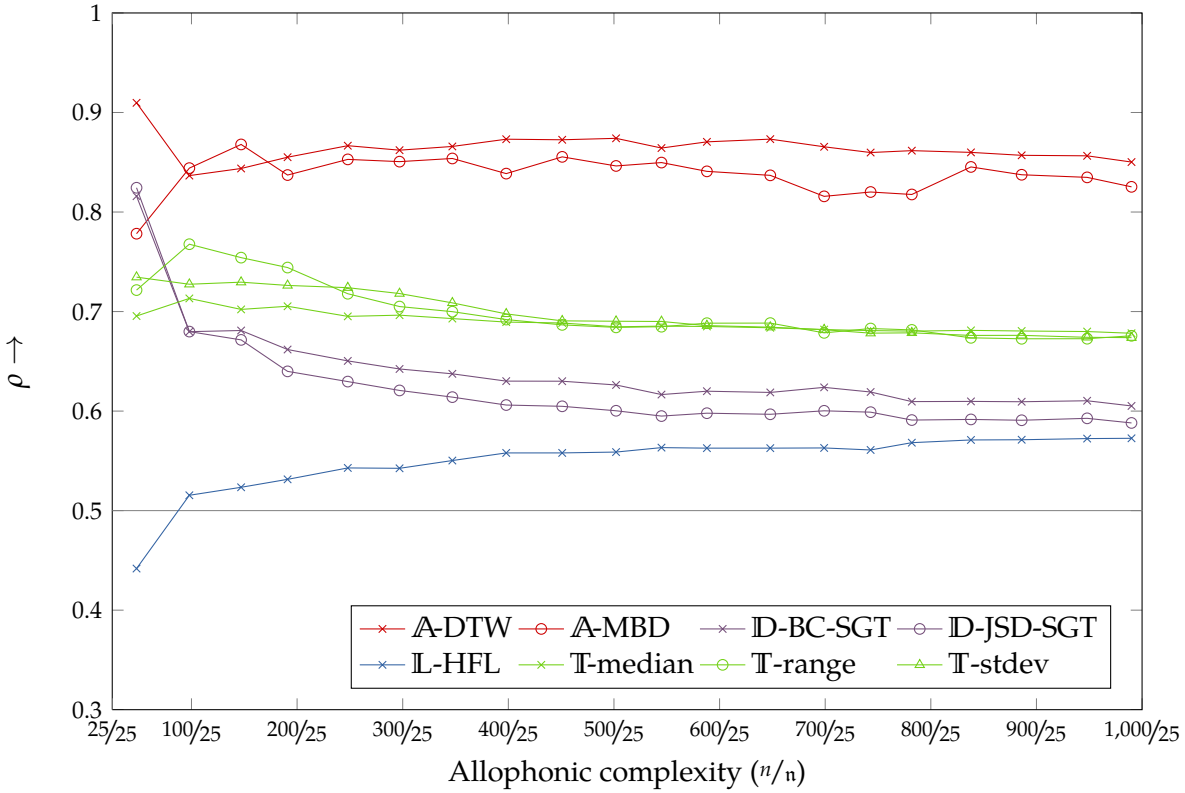


Figure 4.14 — Prognosis of class separation by the ρ statistic for skyline indicators of allophony, as a function of allophonic complexity. The gray line indicates chance.

numerical vs. logical) appeared to outperform individual indicators (Boruta, 2011b). Moreover, these combination schemes were simply ad hoc manipulations of the various values different indicators assign to a given phone pair.

The experiments to be presented in the remainder of the present study address the issue of the combination of different indicators of allophony using well-known statistical models, viz. logistic regression (in Section 4.5 and Chapter 5) and three-way multidimensional scaling (in Chapter 6). Before turning to the actual combination experiments, we need to select which indicators will be combined. Indeed, for each class of indicator, we defined and prognosticated various of them: 2 acoustic indicators, 4 temporal indicators, 5 distributional indicators, and 4 lexical indicators. Accordingly, there are $2 \times 4 \times 5 \times 4 = 160$ possible combinations involving one indicator of each class. However, benchmarking the performance of 160 different combinations for each allophonic complexity is computationally intractable, if relevant at all. In this section, the focus is on selecting the optimal indicator of each of the four classes we consider in the present study.

Take me to your leader As noted in Section 3.2, the number of allophonic rules infants must learn is unknown (if assessable at all; Boruta, 2011b). Therefore, searching for the *optimal* indicator of allophony requires examining the behavior of candidate indicators as the allophonic complexity of their input increases. In the case at hand, such behavior can be though of as, for each indicator, a vector collecting the ρ values observed at each tested allophonic complexity. Assuming higher is better—which is the case when comparing ρ values—finding the optimal indicator thus amounts to finding the maximal vector in a set of vectors. While finding the maximal value in a set of numbers is straightforward, there is no such standard definition for multidimensional data. One answer to the maximal vector problem is the so called skyline query (Papadias et al., 2005; Godfrey et al., 2007) whereby the skyline of a set of vectors contains the vectors that are not dominated by any other vector in the set. Formally, the skyline of a set X is

given by

$$\text{skyline}(X) \equiv \left\{ \mathbf{x} \in X : \nexists\, \mathbf{x}' \in X, \text{dominates}(\mathbf{x}', \mathbf{x}) \right\}. \tag{4.38}$$

A ν-dimensional vector $\mathbf{x} \equiv (x_1, \ldots, x_\nu)$ is said to dominate another vector $\mathbf{x}'$ if each of its components has an equal or better value than the other vector's corresponding component, and if at least one of its components has a strictly better value than the other vector's corresponding component, hence

$$\text{dominates}(\mathbf{x}, \mathbf{x}') \equiv [\![\forall\, v \in (1, 2, \ldots, \nu), x_v \geq x'_v]\!]\, [\![\exists\, v \in (1, 2, \ldots, \nu), x_v > x'_v]\!]. \tag{4.39}$$

As denoted by the set-builder notation in Equation 4.38, the skyline of a set of vectors needs not be a single vector.

Skyline indicators Relying on the prognoses of allophony reported in the previous section, the skyline of each class of indicators are as follows, as illustrated in Figure 4.14:

— 𝔸-DTW and 𝔸-MBD for acoustic indicators;
— 𝕋-median, 𝕋-range, and 𝕋-stdev for temporal indicators;
— 𝔻-BC-SGT and 𝔻-JSD-SGT for distributional indicators;
— 𝕃-HFL for lexical indicators.

Although an objective criterion, the skyline query on ρ values is not sufficient to select a single optimal indicator for each class. Therefore, we decided to reduce the combinatioral possibilities down to 2 indicators per class, i.e. $2^4 = 16$ possible combinations. As skyline queries returned exactly 2 indicators for both acoustic and distributional indicators, 𝔸-DTW, 𝔸-MBD, 𝔻-BC-SGT, and 𝔻-JSD-SGT were maintained without further discussion. As far as temporal indicators are concerned, only 𝕋-median and 𝕋-range were maintained on the grounds that they both rely on robust estimators of central tendency and statistical dispersion, respectively. Finally, the skyline query on lexical indicators only returned 𝕃-HFL; in order to gain further insights on the relative performance of indicators based on functional load and terminimal pairs, we also selected 𝕃-WTP as a candidate indicator.

4.4.3 Confusion plots: a look at indicators' distributions

In order to get a better understanding of the data at hand, we introduce in this section a novel visualization technique, referred to as a confusion plot, that allows for the exploration of the classwise distribution of an indicator's values. In the first section of this chapter, we indeed assumed that the values of an effective indicator of allophony follow a bimodal distribution whose modes emblematize the dichotomy between allophonic pairs and non-allophonic pairs, keeping the ones apart from the others (cf. Assumption 4.1).

Confusion plots The confusion plot for a given indicator at a given allophonic complexity consists of two conjoined, back-to-back histograms, with one representing the distribution of the indicator's values for allophonic pairs and the other representing the corresponding distribution for non-allophonic pairs. In order for the two histograms to be comparable, the same breakpoint values are used to define the cells used on both x-axes, and both y-axes are scaled so that the most populated cell across both histograms accounts for half of the height of the whole plot. Although using a logarithmic scale might prove to be helpful when the data covers a large range of values (cf. Figure 4.4), all confusion plots will be plotted against a linear scale so as not to minimize, in that event, the extreme skewness of the distributions.

A confusion plot is to be interpreted as a (tilted) population pyramid. However, in the case at hand, a pyramidal shape would indicate a poor separation of the two allophonic statuses in the data, as allophonic and non-allophonic pairs would be mainly located in corresponding histogram cells. By contrast, a desirable pattern would consist in a confusion plot where all values assigned to allophonic pairs are massed to the very left of the x-axis (indicating that allophonic pairs tend to be very similar) and all values assigned to non-allophonic pairs are massed to the far right of the x-axis (indicating that non-allophonic pairs tend to be very dissimilar).

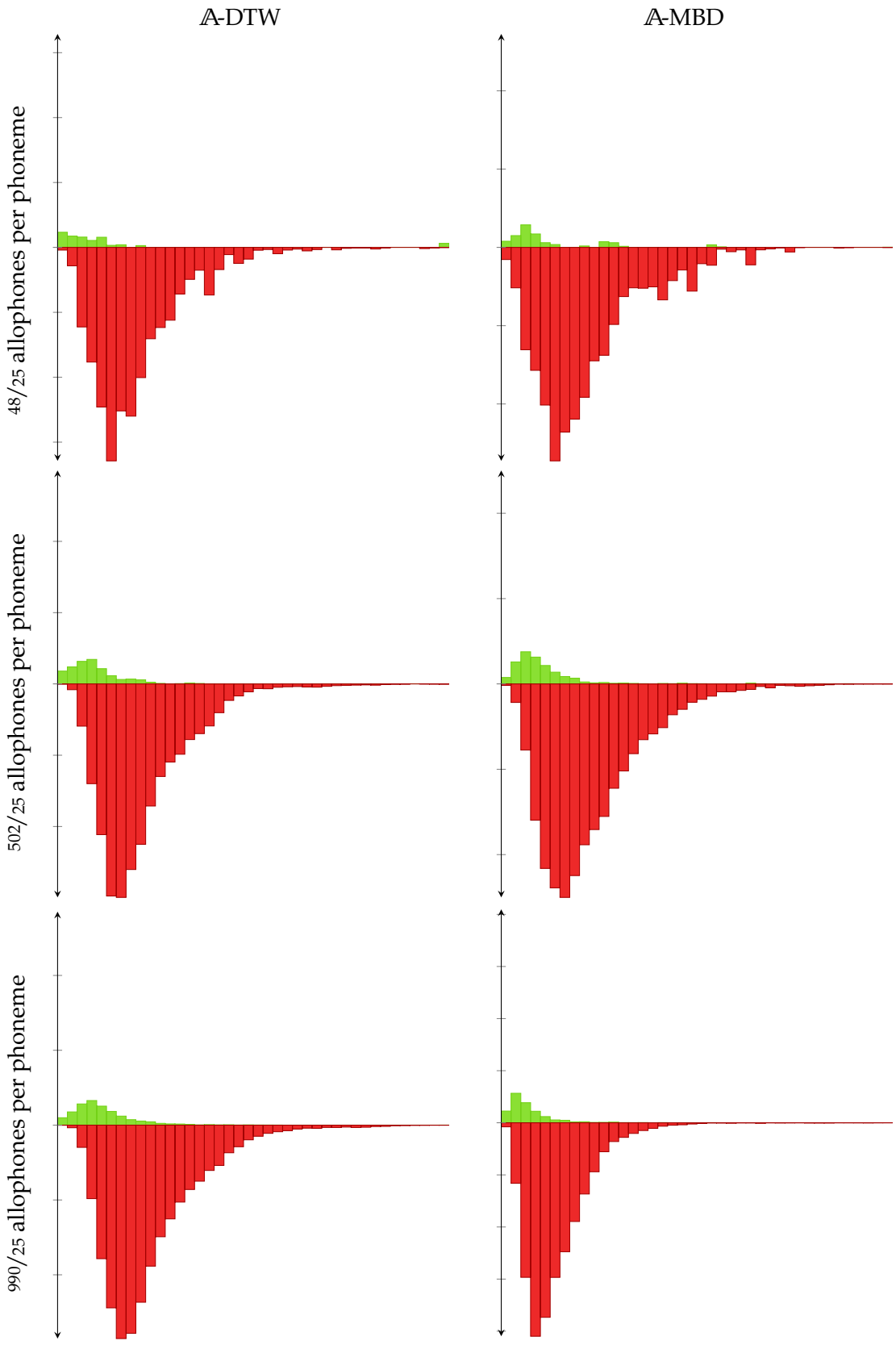


Figure 4.15 — Confusion plots for the acoustic indicators 𝔸-DTW and 𝔸-MBD. For each indicator and allophonic complexity, the upper (green) and lower (red) histograms represent the distribution of this indicator's values at this complexity for allophonic and non-allophonic pairs, respectively. For each plot, the same breakpoint values are used to define the cells used on the x-axes, and both y-axes are scaled so that the most populated cell across both histograms accounts for half of the height of the whole plot.

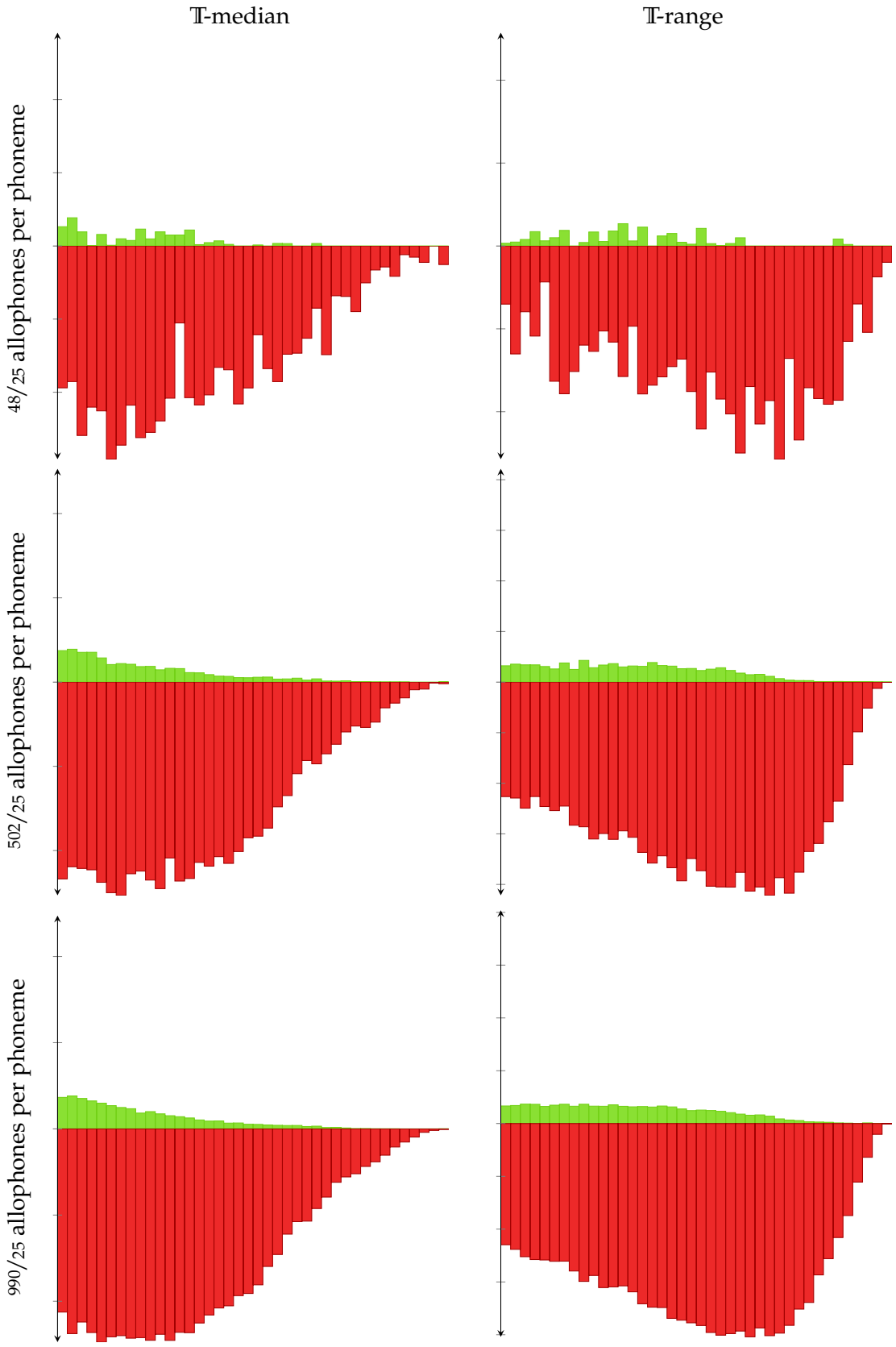


Figure 4.16 — Confusion plots for the temporal indicators 𝕋-median and 𝕋-range. For each indicator and allophonic complexity, the upper (green) and lower (red) histograms represent the distribution of this indicator's values at this complexity for allophonic and non-allophonic pairs, respectively. For each plot, the same breakpoint values are used to define the cells used on the x-axes, and both y-axes are scaled so that the most populated cell across both histograms accounts for half of the height of the whole plot.

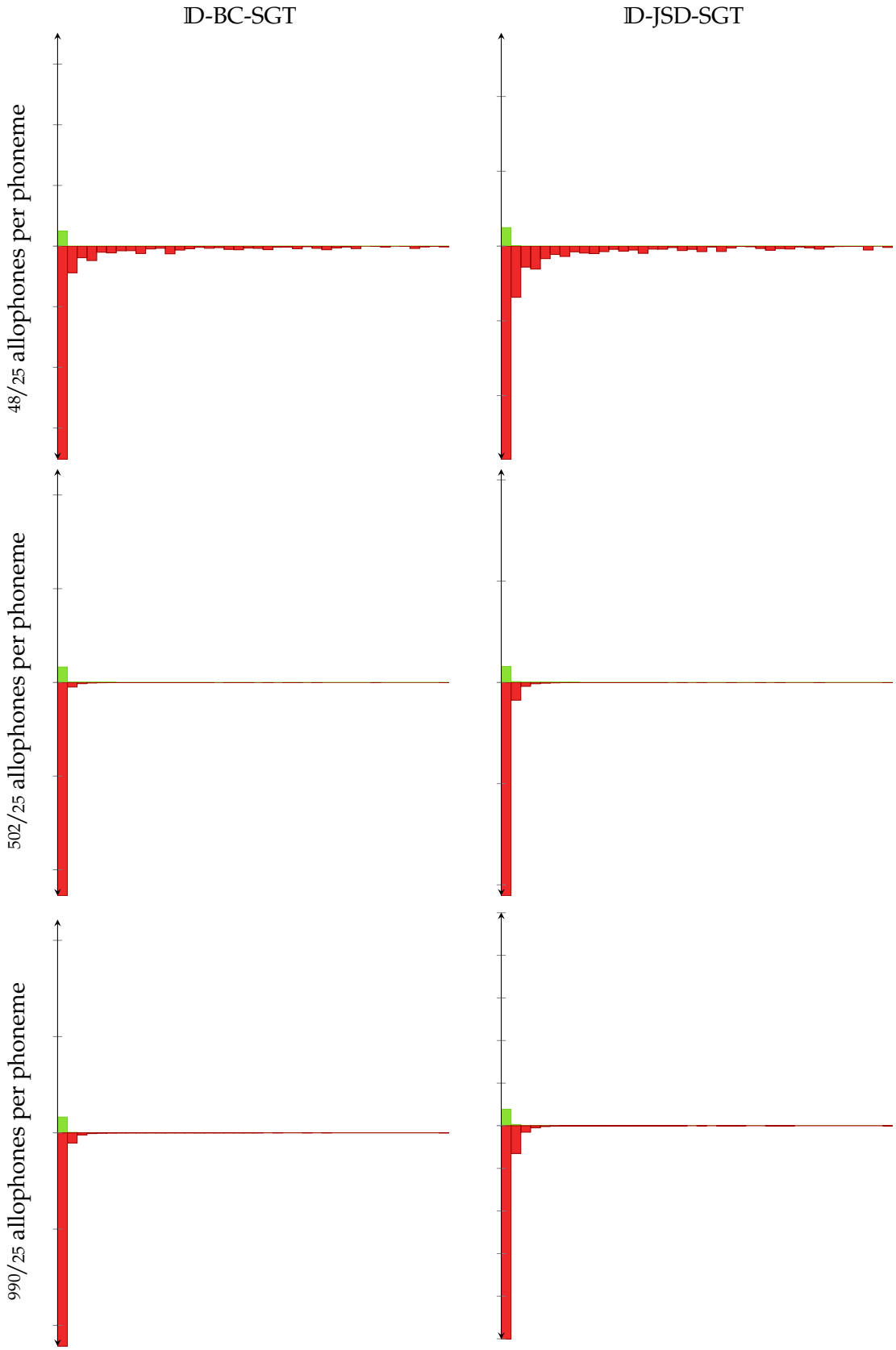


Figure 4.17 — Confusion plots for the distributional indicators 𝔻-BC-SGT and 𝔻-JSD-SGT. For each indicator and allophonic complexity, the upper (green) and lower (red) histograms represent the distribution of this indicator's values at this complexity for allophonic and non-allophonic pairs, respectively. For each plot, the same breakpoint values are used to define the cells used on the x-axes, and both y-axes are scaled so that the most populated cell across both histograms accounts for half of the height of the whole plot.

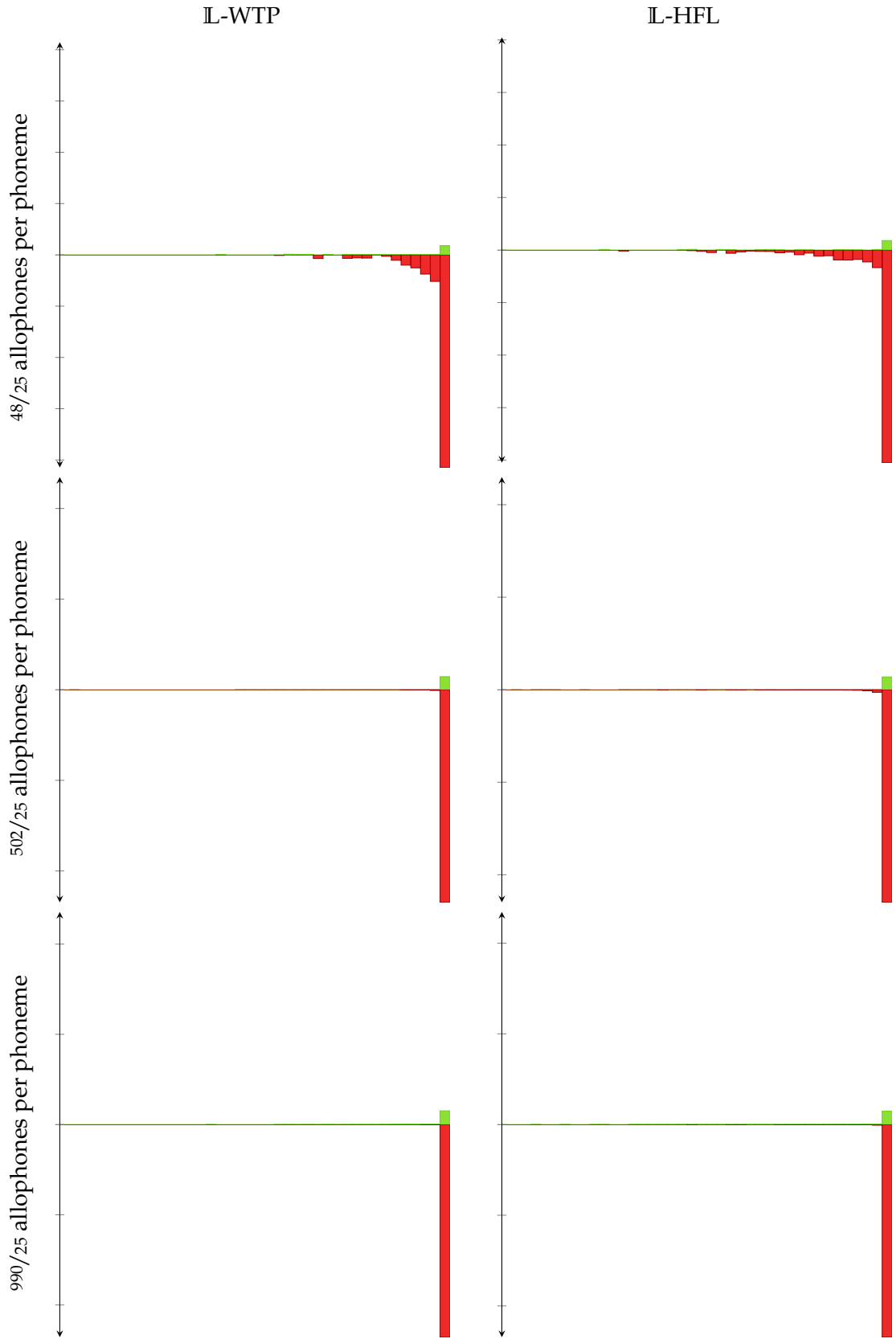


Figure 4.18 — Confusion plots for the lexical indicators 𝕃-WTP and 𝕃-HFL. For each indicator and allophonic complexity, the upper (green) and lower (red) histograms represent the distribution of this indicator's values at this complexity for allophonic and non-allophonic pairs, respectively. For each plot, the same breakpoint values are used to define the cells used on the x-axes, and both y-axes are scaled so that the most populated cell across both histograms accounts for half of the height of the whole plot.

Results For the sake of brevity, we only report confusion plots at low, medium, and high allophonic complexities, i.e. $^{48}/_{25}$, $^{502}/_{25}$, and $^{990}/_{25}$ allophones per phoneme, respectively. These confusion plots are presented in Figures 4.15, 4.16, 4.17, and 4.18 for acoustic, temporal, distributional, and lexical indicators of allophony, respectively.

First and foremost, all confusion plots call for a general observation: the range of values covered by non-allophonic pairs is always larger than the one covered by allophonic pairs. Put another way, non-allophonic pairs are not only approximately 15 times more numerous than allophonic pairs (cf. Figure 4.5), they are also are also much more scattered. Although two modes emerge in both acoustic indicators' confusion plots, as presented in Figure 4.15, it is worth noting that the complete distributions would certainly be unimodal—the bumps emblematizing the distribution of allophonic pairs are only visible because the computation of confusion plots purposedly distinguishes between both allophonic statuses. These observations are confirmed for both temporal indicators' confusion plots, as presented in Figure 4.16; in each plot, the range of values covered by allophonic pairs is fully included in the range covered by non-allophonic pairs. Moreover, in the case of $\mathbb{T}$-median, a pyramidal pattern seems to emerge as the allophonic complexity of the input increases, suggesting that all phone pairs tend to be temporally very similar as the complexity increases. Finally, the confusion plots for distributional and lexical indicators, as presented in Figures 4.17 and 4.18, call for similar observations. First, whereas phone-to-phone dissimilarity values cover a wide range at any allophonic complexity, a significant part of the population is massed in the vicinity of one extreme value: the minimal value for $\mathbb{D}$-BC-SGT and $\mathbb{D}$-JSD-SGT (indicating that all phone pairs, allophonic and non-allophonic, tend to have dissimilar distributions; cf. Figure 4.8), and the maximal value for $\mathbb{L}$-WTP and $\mathbb{L}$-HFL (indicating that all phone pairs tend to be responsible for lexical contrasts; cf. Figure 4.9).

As will be further discussed in Chapter 5, one aforementioned limitation of Peperkamp et al.'s framework is that the learner has to process all possible phone pairs to discover which ones are allophones. However, all aforestated observations are to be interpreted as bad prognoses for the discovery of allophony in a pairwise framework. All ρ values and confusion plots suggest that, at any allophonic complexity, allophonic pairs are indeed both statistically outnumbered and distributionally overwhelmed by non-allophonic pairs. Moreover, non-allophonic pairs are nothing but the spurious byproduct of a combinatorial enumeration as, in the end, only allophonic pairs are useful for the definition of the phonemic inventory.

4.5 Predicting allophony: binary classification task

From this section onwards, we turn to actual (classification- or clustering-based) partitioning experiments. For instance, in this particular section, rather than giving a probabilistic prognosis of the separation of phone pairs, we present experiments in which an explicit threshold is set to tell apart putative allophonic from putative non-allophonic pairs.

Conforming to the recurrent argument that computational models of early language acquisition should perform unsupervised learning (e.g. Brent, 1999; Alishahi, 2011), previous studies modeling the acquisition of allophonic pairs either limited their evaluation to prognoses (Boruta, 2009; Dautriche, 2009; Martin et al., 2009; cf. Table 3.4) or relied on statistical criteria to set a threshold value above or below which phone pairs were classified as allophonic (Peperkamp et al., 2006; Le Calvez, 2007; Le Calvez et al., 2007; Boruta, 2011b). Nonetheless, to our knowledge, no empirical upper bounds on the learnability of allophonic pairs have been reported so far. Put another way, having fitted the parameters of a given model using labeled training data of the form $\langle \delta_{ij}, a^*_{ij} \rangle$, i.e. containing both the observed dissimilarity and the true allophonic status of a given phone pair $\{p_i, p_j\} \subseteq P$, are we able to predict the allophonic status of other (previously unseen) phone pairs? This is the question motivating the experiments we present in this section.

The task Building upon the work of Peperkamp et al. (2006; and subsequent studies), and as illustrated in Figure 4.19, we aim at predicting for any pair of phones $\{p_i, p_j\} \subseteq P$ in a given

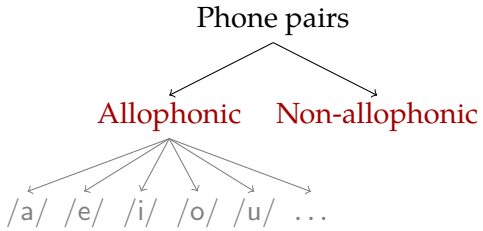


Figure 4.19 — Task diagram: binary allophony on phone pairs. Following Peperkamp et al. (2006), the task consists in predicting whether or not two phones are realizations of the same phoneme; thus, the target categories are the two allophonic statuses (here in red).

allophonic inventory P whether p_i and p_j are allophones, i.e. whether they are realizations of the same underlying phoneme. To this aim, we rely on the aforementioned compactness hypothesis and, consequently, on the definition of indicators of allophony as phone-to-phone dissimilarity measurements. Accordingly, we assumed in Section 4.1 that the conditional probability $\mathrm{P}(a^*_{ij} = 1 \mid \delta_{ij})$ of the phone pair $\{p_i, p_j\}$ being allophonic given the dissimilarity δ_{ij} is inversely proportional to δ_{ij}.

Additionally, let us recall that the task at hand involves predicting one of two mutually exclusive categories (viz. allophony and non-allophony), and that we do not have the same interest in both categories: while allophonic pairs are involved in the definition of phonemes, non-allophonic pairs are nothing but spurious byproducts of a combinatorial enumeration. In that sense, we can consider allophonic pairs to be successes, and non-allophonic pairs to be failures.

4.5.1 Binomial logistic regression

Binomial logistic regression (Faraway, 2005, 2006; Agresti, 2007; Chen et al., 2012) is a classic statistical technique allowing for the modeling of such binary outcomes. More precisely, binomial logistic regression is a type of regression analysis used to model and predict the outcomes of a binary random variable (in our case, the phone pairs' allophonic statuses) based on an arbitrary number of predictor variables (the various indicators of allophony). Let

$$\pi_{ij} \equiv \mathrm{P}(a^*_{ij} = 1 \mid \delta_{ij1}, \ldots, \delta_{ij\kappa}) \tag{4.40}$$

denote the conditional probability of the phone pair $\{p_i, p_j\} \subseteq P$ being allophonic given the κ predictor values $\delta_{ij1}, \ldots, \delta_{ij\kappa}$. The logistic regression model has linear form for the logit of the probability of allophony, i.e.

$$\mathrm{logit}(\pi_{ij}) \equiv \alpha + \sum_{k=1}^{\kappa} \beta_k \delta_{ijk} \qquad \text{where} \qquad \mathrm{logit}(\pi) \equiv \log\Big(\frac{\pi}{1-\pi}\Big), \tag{4.41}$$

and, hence

$$\pi_{ij} = \mathrm{expit}\Big(\alpha + \sum_{k=1}^{\kappa} \beta_k \delta_{ijk}\Big) \qquad \text{where} \qquad \mathrm{expit}(x) \equiv \frac{\exp(x)}{1+\exp(x)}. \tag{4.42}$$

As illustrated in Figure 4.20, the logit function defines an S-shaped link between the weighted linear combination of the predictor values and the probability of allophony. Because the regression weights $\beta_1, \ldots, \beta_\kappa$ and the intercept α can be either positive or negative—albeit valid, a null coefficient would merely cancel out the corresponding predictor—binomial logistic regression models can indifferently accomodate predictors that are positively or negatively correlated with the response variable. In any case, the estimated probability of success increases as the weighted combination of the predictor values increases.

Underlying assumptions It is worth noting that logistic regression models make no assumption on the distribution of the predictor variables; they do not need, for example, to be normally distributed or linearly related. Considering the rather shapeless distributions we observed in the

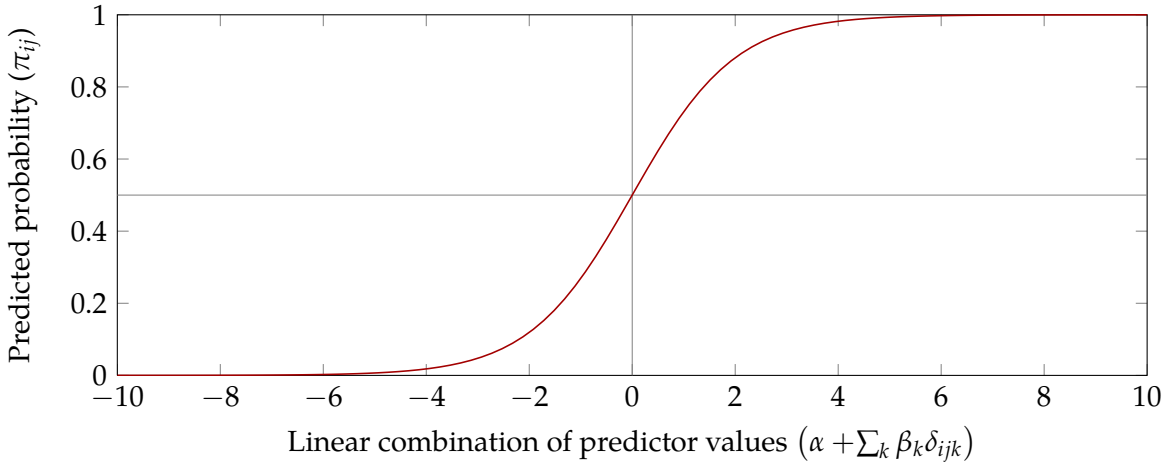


Figure 4.20 — The logistic function.

confusion plots in Section 4.4.3, this insensitivity to the shape of the distribution of the predictor variables justifies the use of logistic regression in the case at hand. Nonetheless, using binomial logistic regression entails the following assumptions (Agresti, 2007; Chen et al., 2012):

Assumption 4.2 The true outcome probabilities follow a binomial distribution. Thereby, the observations are independent, meaning that the allophonic status of a given pair of phones makes it neither more nor less probable that another one is allophonic or not.

Assumption 4.3 The true outcome probabilities are a logistic function of a (weighted) linear combination of the predictor variables.

Assumption 4.4 No important predictor variables are omitted, and no extraneous predictor variables are included, meaning that the indicators used in the model are individually necessary and collectively sufficient.

Assumption 4.5 The predictor variables are not linear combinations of each other, meaning that no two indicators yield identical rankings of the phone pairs.

Assumption 4.6 The predictor variables are measured without error, meaning that indicators are estimated on a representative sample, and using robust methods.

In light of the substantial size of the CSJ, we consider this last assumption to be fulfilled without further discussion. Furthermore, Assumption 4.5 appears to be satisfied, too, as no two indicators have identical ρ values at any allophonic complexity (cf. Section 4.4.1). The first three assumptions, however, are non-trivial in the case at hand. Although we argued that all aforementioned indicator classes are individually necessary, Assumption 4.4 entails that indicators should be collectively sufficient: except from a post hoc, empirical validation (i.e. if all predictions are successful), we can not guarantee that this assumption be fulfilled. The same reasoning holds for Assumption 4.3: to us, the main attraction of binomial logistic regression arises from its built-in ability to combine indicators and assessing whether the logistic function is the most appropriate link function is out of the stated scope of this research.

Finally, although already implicitly present in all experiments building upon Peperkamp et al.'s (2006) study, we know for sure that Assumption 4.2 is not satisfied in the case at hand. Indeed, statistical independence can not, by definition, be assumed for phone pairs because phone pairs are not atomic objects (cf. Section 4.1) and knowing, for example, that both phone pairs $\{p_i, p_j\}$ and $\{p_{i'}, p_j\}$ are allophonic should by transitivity entail the prediction that the pair $\{p_i, p_{i'}\}$ is allophonic. For the sake of comparability with previously reported studies on the acquisition of allophonic pairs, we will consider Assumption 4.2 to be a necessary simplifying assumption.

Model-fitting In the case of binomial logistic regression, fitting the model consists in estimating the regression parameters $\beta_1, \ldots, \beta_\kappa$ and the intercept α from a given a set of training observations of the form $\langle \delta_{ij1}, \ldots, \delta_{ij\kappa}, w_{ij}, a^*_{ij} \rangle$, i.e. comprising the dissimilarity ratings assigned to a given phone pair $\{p_i, p_j\}$, its weight, and its true allophonic status. Exposing the details of the estimation process is beyond the scope of the present study. Suffice it to say that the numerical algorithm we used outputs the parameter values that maximize the log-likelihood of the parameters given the training data (Agresti, 2007; R Development Core Team, 2010; cf. `stats::glm`). Adapting the definition given by Faraway (2006; pp. 30, 129) to our notation, the log-likelihood function is given by

$$\text{log-likelihood}(\alpha, \beta_1, \ldots, \beta_\kappa; \mathbf{A}^*, \mathbf{W}) \equiv \sum_{i<j} a^*_{ij} \operatorname{logit}(\pi_{ij}) - w_{ij} \operatorname{logit}(\pi_{ij}) + \log \binom{w_{ij}}{a^*_{ij} w_{ij}}. \quad (4.43)$$

As pointed out by Agresti (2007), a desirable property of regression models is that they smooth the sample data, somewhat dampening the observed variability.

It is worth mentioning that if any predictor variable indicates a complete separation of the two classes (i.e. if there is a value at and above which all observations are successes and below which all observations are failures, or vice versa), the estimation of the parameters by maximum likelihood is made impossible (Chen et al., 2012). In the case at hand, however, we already observed that no predictor contains such ideal data. Indeed, the probabilistic prognosis of allophony we presented in Section 4.4.1 is nothing but a numerical assessement of class separation, and for no indicator did we observe ρ values equal to 1, at any allophonic complexity.

Individual and quartet models Because, to our knowledge, these experiments are the first attempt at using binomial logistic regression to model the acquisition of allophony, and because many indicators of allophony were introduced in the present study, we compare the performance of regression models relying on a single indicator (i.e. $\kappa = 1$, referred to as individual models) to that of models relying on a combination comprising one indicator of each class (i.e. $\kappa = 4$, referred to as quartet models). As we reduced the number of indicators of allophony to two indicators per class, viz.

$$\left\{\begin{matrix}\mathbb{A}\text{-DTW} \\ \mathbb{A}\text{-MBD}\end{matrix}\right\} \times \left\{\begin{matrix}\mathbb{T}\text{-median} \\ \mathbb{T}\text{-range}\end{matrix}\right\} \times \left\{\begin{matrix}\mathbb{D}\text{-JSD-SGT} \\ \mathbb{D}\text{-BC-SGT}\end{matrix}\right\} \times \left\{\begin{matrix}\mathbb{L}\text{-WTP} \\ \mathbb{L}\text{-HFL}\end{matrix}\right\},$$

there are $2 \times 4 = 8$ possible individual models and, consequently, $2^4 = 16$ possible quartet models. Unless otherwise specified, logistic regression models rely on the aforementioned range standardization of the indicators. However, as discussed in Section 4.3.2, these experiments address an additional, transversal issue, viz. deciding whether allophony can be learned from a mere ranking of an indicator's values or if the relative values bring additional information. For this reason, each of the $8 + 16 = 24$ regression models will be evaluated on both the range- and the rank-based standardization of the indicators' data.

4.5.2 Evaluation

In these experiments, we are not interested in the interpretation of the estimated regression parameters, but in the predictive power of the regression models. For the purpose of classifying candidate phone pairs as allophonic or non-allophonic, we rely on the standard threshold of $\pi_{ij} > .5$ whereby a given phone pair $\{p_i, p_j\}$ is considered to be allophonic, i.e. when

$$\mathrm{P}(a^*_{ij} = 1 \mid \delta_{ij1}, \ldots, \delta_{ij\kappa}) > \mathrm{P}(a^*_{ij} = 0 \mid \delta_{ij1}, \ldots, \delta_{ij\kappa}). \quad (4.44)$$

As a matter of convenience, let $\mathbf{A} \equiv [a_{ij}]$ denote the predicted matrix of allophony where $a_{ij} \equiv [\![\pi_{ij} > .5]\!]$ denotes the predicted allophonic status of the pair $\{p_i, p_j\}$.

Confirmatory overfitting Generally speaking, predictive models are trained and tested on distinct sets of observations so as to avoid overfitting, and binomial logistic regression models

are no exception. Overfitting can be defined as the discovery of apparent predictive relations in the training data (i.e. between the predictor variables $\delta_{ij1}, \ldots, \delta_{ij\kappa}$ and the true outcome a^*_{ij}) that do not hold in general. Using distinct yet comparable observation sets allows for an assessment of the overfitting of the model parameters to the training data: if a given model fits both the training and the test data, it is unlikely that overfitting occurred; by contrast, if a model fits the training data much better than the test data, then one can suspect the model to be overfitted.

However, in the case at hand, we aim at providing empirical upper bounds on the learnability of allophony and, therefore, chose to perform supervised learning to place ourselves in a best-case scenario. To further help the discovery of the dichotomy between allophonic and non-allophonic pairs, we purposedly use all available data for both the training and the evaluation of our regression models, a setup we tentatively refer to as confirmatory overfitting. Put another way, given all available data, and including the reference matrix of allophony $\mathbf{A}^*$, can we tell apart allophonic and non-allophonic pairs?

It is worth noting that we did perform more reasonable tests in preliminary experiments. Indeed, for all individual and quartet models, we compared their predictive power using confirmatory overfitting and two standard techniques of cross-validation. First, we carried out evaluations using a technique known as tenfold cross-validation, whereby all available observations are partitioned into 10 subsamples, 1 subsample is reserved as the test set and the remaining 9 are used as the training set; this process being repeated 10 times so that each subsample is used exactly once as the test set. Second, we carried out evaluations using a cross-validation technique known as leave-one-out validation, whereby a single observation is reserved as the test set and all other observations are used as the training data; this process being repeated so that each each observation is used exactly once as the test set. Because these techniques require that a considerable number of regression models be trained—10 and $n(n-1)/2$ for tenfold and leave-one-out cross-validation, respectively, for each of the 24 models and the 20 tested allophonic complexities—they are computationally prohibitive. Accordingly, these preliminary experiments were only carried out for the lowest allophonic complexities, viz. $^{48}/_{25}$, $^{98}/_{25}$, and $^{147}/_{25}$ allophones per phoneme. For all benchmarked models and complexities, the difference in performance between confirmatory overfitting and both cross-validations techniques did not exceed $\pm 2\%$ of accuracy, thus indicating that confirmatory overfitting does not have a significant impact on the predictive power of the models.

Goodness of fit The first quantitative criterion we use to assess the fit of a binomial logistic regression model to the reference outcomes collected in $\mathbf{A}^*$ is known as the Akaike information criterion (henceforth AIC; Akaike, 1974). The major piece of information the AIC relies on is the previously defined likelihood function (cf. Equation 4.43). Indeed, the AIC of a given binomial logistic regression model is given by the maximized value of the likelihood function for its parameters, penalized by the number of such parameters. Concretely, it is defined as

$$\mathrm{AIC}(\alpha, \beta_1, \ldots, \beta_\kappa; \mathbf{A}^*, \mathbf{W}) \equiv -2\big(\text{log-likelihood}(\alpha, \beta_1, \ldots, \beta_\kappa; \mathbf{A}^*, \mathbf{W}) - (\kappa + 1)\big) \tag{4.45}$$

where $\beta_1, \ldots, \beta_\kappa$ and α denote the estimated optimal paramater values. It is worth noting that the last term $\kappa + 1$ matches the number of such parameters in the model (κ distinct β_k coefficients, plus one α coefficient). The penalty accounts for the fact that increasing the number of parameters (i.e. of predictor variables) tends to increase the goodness of fit, regardless of the true number of parameters that may sufficiently account for the observations. Regression models with a good fit tend to minimize the AIC (i.e. the lower the better; Faraway, 2005).

Because the AIC does not rely on a partition of, for a given allophonic inventory, all possible phone pairs into allophonic and non-allophonic pairs, and following the discussion in Section 4.4, this criterion is to be considered as a prognosis rather than as a true evaluation measure.

Contingency table The other quantitative criteria we use to assess the performance of binomial logistic regression models rely on the models' predictive power. For each model, a *contingency table* (a.k.a. confusion matrix or classification table) is used to cross-classify the true outcomes

Table 4.2 — Schematic 2×2 contingency table for binomial logistic regression.

	$a_{ij} = 1$	$a_{ij} = 0$
$a^*_{ij} = 1$	*TP*	*FN*
$a^*_{ij} = 0$	*FP*	*TN*

with the model's predictions. In the binary case of allophony, such a cross-classification yields a 2×2 contingency table, as illustrated in Table 4.2. The non-negative counts in each cell are to be interpreted as follows.

A true positive (henceforth *TP*, a.k.a. hit) occurs when an allophonic pair is indeed classified as allophonic. Similarly, a true negative (henceforth *TN*, a.k.a. correct rejection) occurs when a non-allophonic pair is classified as non-allophonic. There are two types of errors we can commit when a prediction does not match the true allophonic status of the pair. A false positive (henceforth *FP*, a.k.a. false alarm and type I error) occurs when a non-allophonic pair is wrongfully classified as allophonic. Conversely, a false negative (henceforth *FN*, a.k.a. miss and type II error) occurs when an allophonic pair is classified non-allophonic. Formally, using the pairwise weights $\mathbf{W}$, the reference allophonic statuses $\mathbf{A}^*$, and the predicted statuses $\mathbf{A}$, these quantities are given by

$$TP \equiv \sum_{i<j} w_{ij} [\![a_{ij} = 1]\!] [\![a^*_{ij} = 1]\!], \tag{4.46}$$

$$FP \equiv \sum_{i<j} w_{ij} [\![a_{ij} = 1]\!] [\![a^*_{ij} = 0]\!], \tag{4.47}$$

$$FN \equiv \sum_{i<j} w_{ij} [\![a_{ij} = 0]\!] [\![a^*_{ij} = 1]\!], \tag{4.48}$$

$$TN \equiv \sum_{i<j} w_{ij} [\![a_{ij} = 0]\!] [\![a^*_{ij} = 0]\!]. \tag{4.49}$$

Based on these four quantities, various evaluation measures can be further computed, as summarized in Table 4.3.

Accuracy A simple and (thus) ubiquitous evaluation measure in the fields of machine learning and computational linguistics, the accuracy (a.k.a. Rand index) of a model is simply defined as the proportion of correct predictions (both *TP* and *TN*) across all predictions. Accuracy values can thus be interpreted as the probability $\mathrm{P}(a_{ij} = a^*_{ij})$ of the predicted allophonic status of a randomly-drawn phone pair $\{p_i, p_j\} \subseteq P$ being identical to the true allophonic status of the pair. Formally, accuracy is given by

$$\text{accuracy}(\mathbf{A}, \mathbf{A}^*) \equiv \mathrm{P}(a_{ij} = a^*_{ij}) \approxeq \frac{\sum_{i<j} w_{ij} [\![a_{ij} = a^*_{ij}]\!]}{\sum_{i<j} w_{ij}} = \frac{TP + TN}{FP + TP + TN + FN}. \tag{4.50}$$

However, as we argued in a previous study (Boruta, 2011b), one limitation of accuracy arises from the fact that it does not distinguish between allophonic and non-allophonic pairs: in the numerators in Equation 4.50, a correct prediction is a correct prediction, regardless if it is a *TP* or a *TN*. Following our argument that non-allophonic pairs are nothing but the byproduct of a combinatorial enumeration as well as our prior observations that allophonic pairs are, regardless of the allophonic complexity of the input, outnumbered by non-allophonic pairs (cf. Table 4.5 and Section 4.4.3), accuracy values are not sufficient to properly understand of the performance of a model on the separation of the two allophonic statuses.

Precision and recall As they do not take the number of *TN* into account, precision and recall appear to be relevant evaluation measures in order to assess the separation of classes in a model's predictions. Following the probabilistic interpretation discussed by Goutte & Gaussier (2005) and, to a lesser extent, Sing et al. (2005), a precision measure can be interpreted as an estimate

Table 4.3 — Summary of the various prognosis and evaluation measures used for the binary classification task. When applicable, an abridged probabilistic interpretation is given as a function of a randomly-drawn phone pair $\{p_i, p_j\} \subseteq P$.

Measure	Threshold	Objective	Range	Probabilistic interpretation
ρ	Not applicable	Maximize	$[0,1]$	$\mathrm{P}(\pi_{ij} > \pi_{i'j'} \mid a^*_{ij} = 0, a^*_{i'j'} = 1)$
AIC	Not applicable	Minimize	$[2(\kappa+1), \infty[$	Not applicable
Accuracy	$\pi_{ij} > 1 - \pi_{ij}$	Maximize	$[0,1]$	$\mathrm{P}(a_{ij} = a^*_{ij})$
Precision	$\pi_{ij} > 1 - \pi_{ij}$	Maximize	$[0,1]$	$\mathrm{P}(a_{ij} = 1 \mid a^*_{ij} = 1)$
Recall	$\pi_{ij} > 1 - \pi_{ij}$	Maximize	$[0,1]$	$\mathrm{P}(a^*_{ij} = 1 \mid a_{ij} = 1)$
F-score	$\pi_{ij} > 1 - \pi_{ij}$	Maximize	$[0,1]$	None
MCC	$\pi_{ij} > 1 - \pi_{ij}$	Maximize	$[-1,1]$	Not applicable

of the probability $\mathrm{P}(a_{ij} = 1 \mid a^*_{ij} = 1)$ of a randomly-drawn allophonic pair $\{p_i, p_j\} \subseteq P$ being classified as allophonic. Formally, precision is given by

$$\mathrm{precision}(\mathbf{A}, \mathbf{A}^*) \equiv \mathrm{P}(a_{ij} = 1 \mid a^*_{ij} = 1) \approxeq \frac{\sum_{i<j} w_{ij} a_{ij} a^*_{ij}}{\sum_{i<j} w_{ij} a^*_{ij}} = \frac{TP}{TP + FP}. \tag{4.51}$$

Defined as an estimate of the probability $\mathrm{P}(a^*_{ij} = 1 \mid a_{ij} = 1)$ of a randomly-drawn phone pair $\{p_i, p_j\} \subseteq P$ classified as allophonic being a true allophonic pair, recall can then be considered to be the corollary of precision. Formally, recall is given by

$$\mathrm{recall}(\mathbf{A}, \mathbf{A}^*) \equiv \mathrm{P}(a^*_{ij} = 1 \mid a_{ij} = 1) \approxeq \frac{\sum_{i<j} w_{ij} a_{ij} a^*_{ij}}{\sum_{i<j} w_{ij} a_{ij}} = \frac{TP}{TP + FN}. \tag{4.52}$$

Finally, both precision and recall can be combined into a composite evaluation measure known as F-score is a composite measure, given by the harmonic mean of the two, i.e.

$$\text{F-score}(\mathbf{A}, \mathbf{A}^*) \equiv \frac{2 \cdot \mathrm{precision}(\mathbf{A}, \mathbf{A}^*) \cdot \mathrm{recall}(\mathbf{A}, \mathbf{A}^*)}{\mathrm{precision}(\mathbf{A}, \mathbf{A}^*) + \mathrm{recall}(\mathbf{A}, \mathbf{A}^*)} = \frac{2 \cdot TP}{2 \cdot TP + FN + FP}. \tag{4.53}$$

As for precision and recall, F-score values lie in $[0,1]$, too. However, to our knowledge, this particular evaluation measure has no probabilistic interpretation.

Matthew's coefficient of correlation Whereas precision, recall, and F-score ignore the number of *TN* and, thus, can be used to asses the separation of classes in a given model's predictions, the fact that not all cells of the contingency table are used in the computation of this measures limits their relevance. Indeed, an artificial learner à la Peperkamp et al. (2006) needs to process all possible pairs of phones in its input and, hence, should be rewarded for the correct rejections of non-allophonic pairs. Therefore, following the protocol we used in a previous study (Boruta, 2011b), the last evaluation measure we consider for the binary task of allophony is Matthew's coefficient of correlation (henceforth MCC; Matthews, 1975) as it is indeed considered to yield a balanced evaluation measure even when both classes are not equally represented in the input. MCC measures the correlation between the true and the predicted allophonic statuses. Formally, it is given by

$$\mathrm{MCC}(\mathbf{A}, \mathbf{A}^*) \equiv \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{4.54}$$

MCC values lie in $[-1,1]$: a coefficient of 1 indicates a perfect prediction, a coefficient of -1 indicates a completely inverse prediction, and 0 indicates that the predictions are no better than random predictions.

Summarizing predictive power Finally, it is worth noting that the aforementioned ρ statistic could also be used to assess the predictive power of a binomial logistic regression model (a.k.a. the concordance index in this context Agresti, 2007; pp. 143–144). Indeed, collecting all predicted

probabilities of non-allophony (i.e. $1 - \pi_{ij}$) in an $n \times n$ matrix $\mathbf{\Pi} \equiv [1 - \pi_{ij}]$, $\rho(\mathbf{\Pi}; \mathbf{A}^*, \mathbf{W})$ is equal to the MLE of the probability of the predicted probability of a randomly-drawn allophonic pair being allophonic being higher than the predicted probability of a randomly-drawn non-allophonic pair being allophonic.

Nonetheless, due to the considerable number of parameters we want to assess (8 individual regression models, 16 quartet regression models, 20 allophonic complexities, and 2 standardizations techniques), the time complexity in $\mathcal{O}(n^2 \log_2 n)$ of the computation of ρ prohibits its use for the $(8 + 16) \times 20 \times 2 = 960$ binomial logistic regression models considered in this experiment.

4.5.3 Results

The performance of all aforementioned binomial logistic regression models is presented in Figures 4.21, 4.22, 4.23, 4.24, 4.25, and 4.26 in terms of AIC, accuracy, precision, recall, F-score and MCC, respectively.

Goodness of fit Let us first consider the goodness of fit tests reported in Figure 4.21. The main observation is that all regression models' AIC increases as the allophonic complexity of their input increases, meaning that the more phonemes have allophones, the less indicators of allophony (or a logistic combination thereof) fit the binary partition of all possible phone pairs into allophonic and non-allophonic pairs. Furthermore, whereas temporal, distributional, and lexical indicators appear to yield the models that least fit the data at hand, acoustic indicators yield models with a significantly better fit, hence confirming the prognosis discussed in Section 4.4.1.

More than that, not only do the AICs of the quartet models suggest that their goodness of fit is approximately bounded by that of the acoustic indicators 𝔸-DTW and 𝔸-MBD, they also show that the best quartet models have a better goodness of fit than 𝔸-DTW. The reason for this may be twofold: combining indicators of allophony of different classes yields models that better fit the data or—in spite of the penalty in the definition of the AIC (cf. Equation 4.45)—including more predictor variables merely yields overfitted regression models. Nonetheless, AIC is an unbounded criterion and, hence, we can not formulate but comments comparing the AIC of a given model to the AIC of another model.

Classification performance By contrast, all the evaluation measures we consider in the present study are bounded. For instance, according to the accuracy measures presented in Figure 4.22, binomial logistic regression models appear to be very effective at distinguishing between allophonic and non-allophonic pairs in their input: indeed, even at maximal allophonic complexity, all (individual and quartet) regression models classified their input with an accuracy strictly greater than 93%. Moreover, it is worth emphasizing that whereas accuracy values lie in $[0, 1]$, the y-axis in Figure 4.22 was scaled to the range $[.93, 1]$ for the sake of readability, hence exaggerating the models' sensitivity to an increase in allophonic complexity—all models' accuracy plateaus from $^{450}/_{25}$ allophones per phoneme onwards. Furthermore, as far as classification accuracy is concerned, quartet models are (on average) at least as good as 𝔸-MBD, thus confirming our prior observation that combining indicators increases the quality of the predictions in this binary task.

Surprisingly, accuracy values in Figure 4.22 suggest that lexical indicators are more effective than temporal or distributional indicators for the purpose of telling apart allophonic from non-allophonic pairs. Although this contradicts the prognosis of allophony reported in Section 4.4.1, this observation is corroborated by the precision and recall measures reported in Figures 4.23 and 4.24 whereby—regardless of the allophonic complexity of their input—all individual models trained with temporal or distributional indicators have strictly null performance. The interpretation of such values is however straightforward: a binomial logistic regression model whose precision and recall are both equal to 0, but whose accuracy is strictly positive actually classified all observations in its input as failures—i.e. non-allophonic pairs in the case at hand. Indeed, whereas precision and recall focus on the number of *TP*, accuracy benefits from the number of

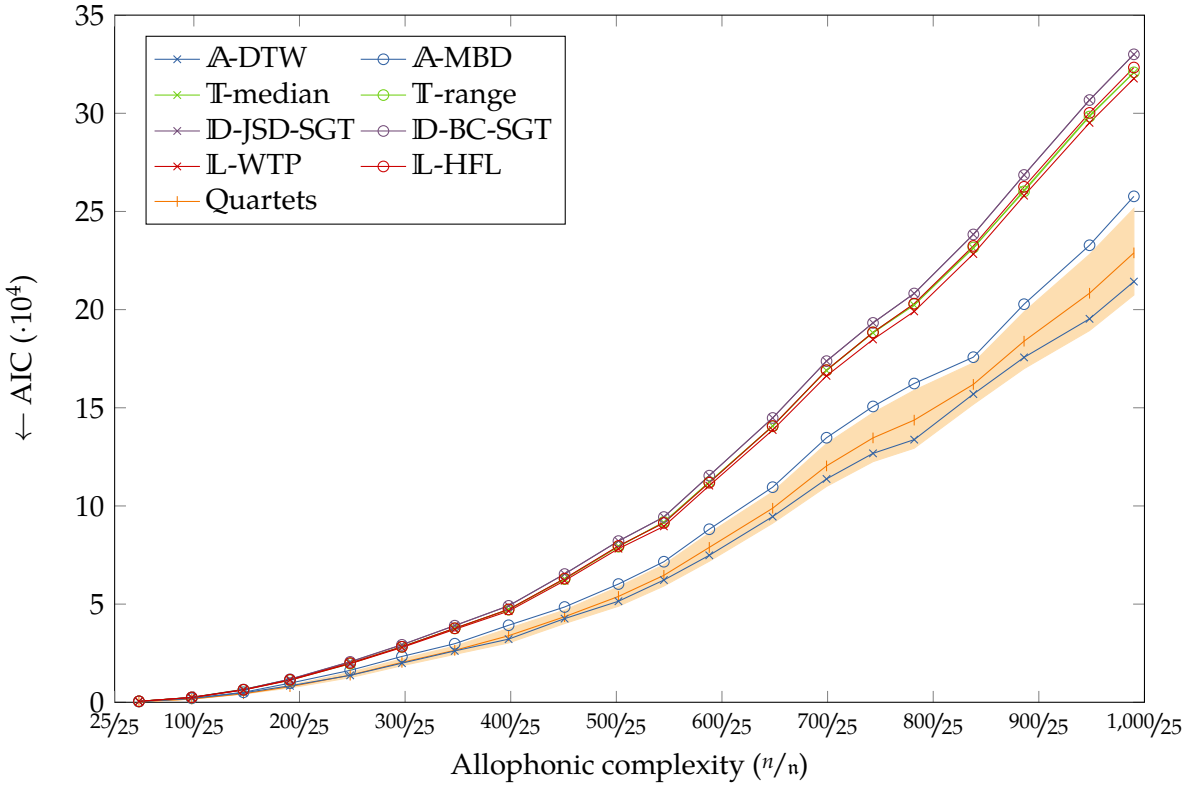


Figure 4.21 — AIC of the logistic regression models on the binary classification task. For quartet models, the solid line marks the average AIC, and the band is bounded by the lowest and the highest values for each allophonic complexity.

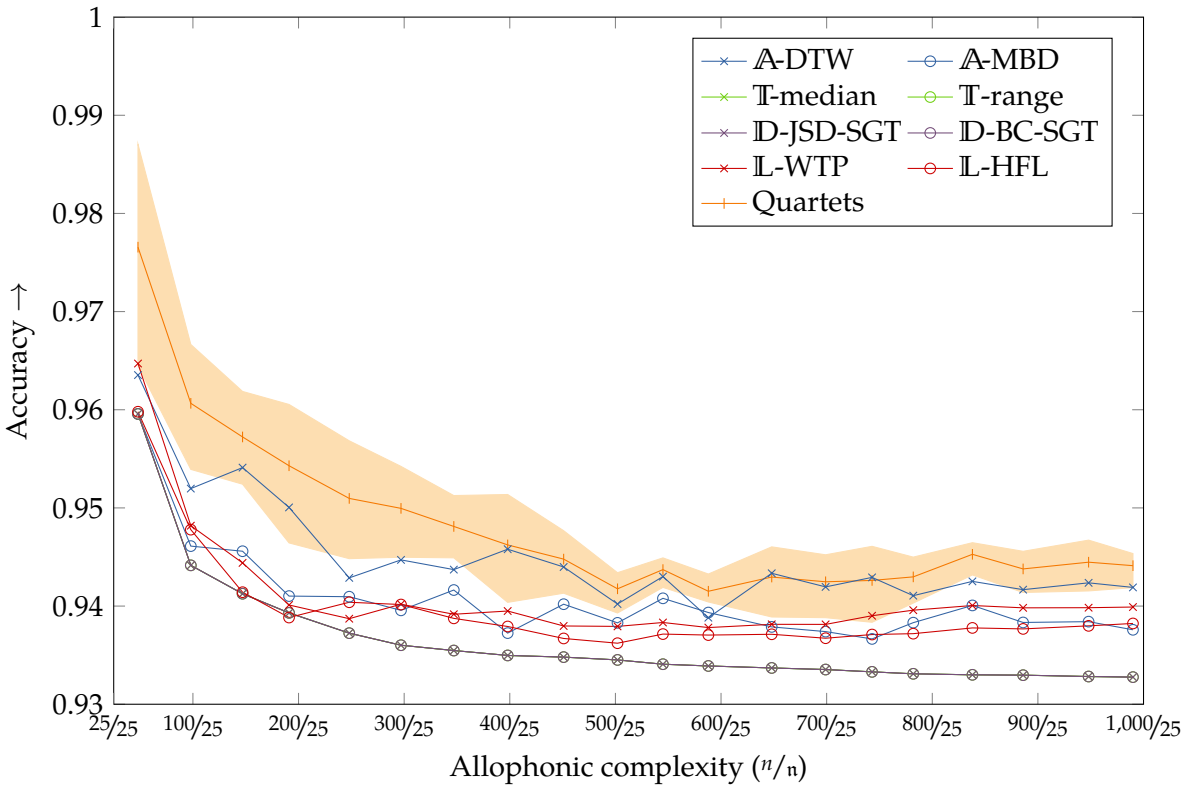


Figure 4.22 — Accuracy of the logistic regression models on the binary classification task. For quartet models, the solid line marks the average accuracy, and the band is bounded by the lowest and the highest values for each allophonic complexity.

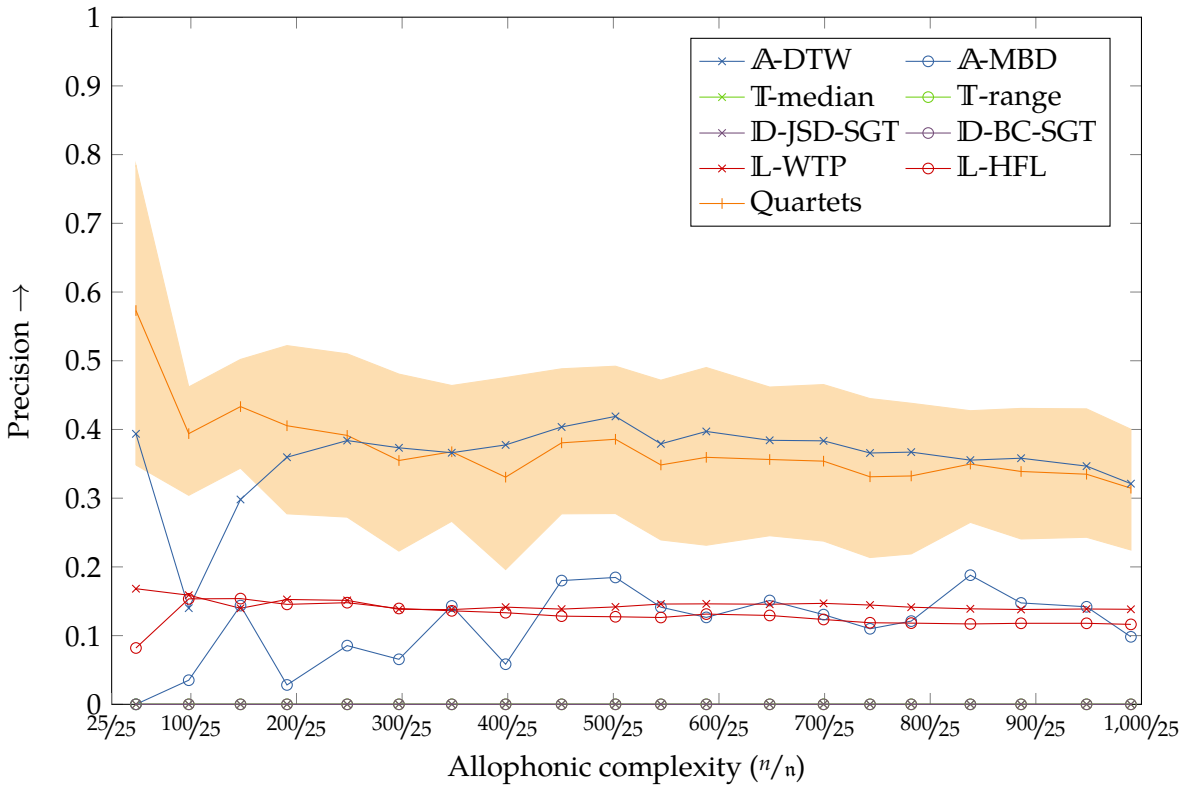


Figure 4.23 — Precision of the logistic regression models on the binary classification task. For quartet models, the solid line marks the average precision, and the band is bounded by the lowest and the highest values for each allophonic complexity.

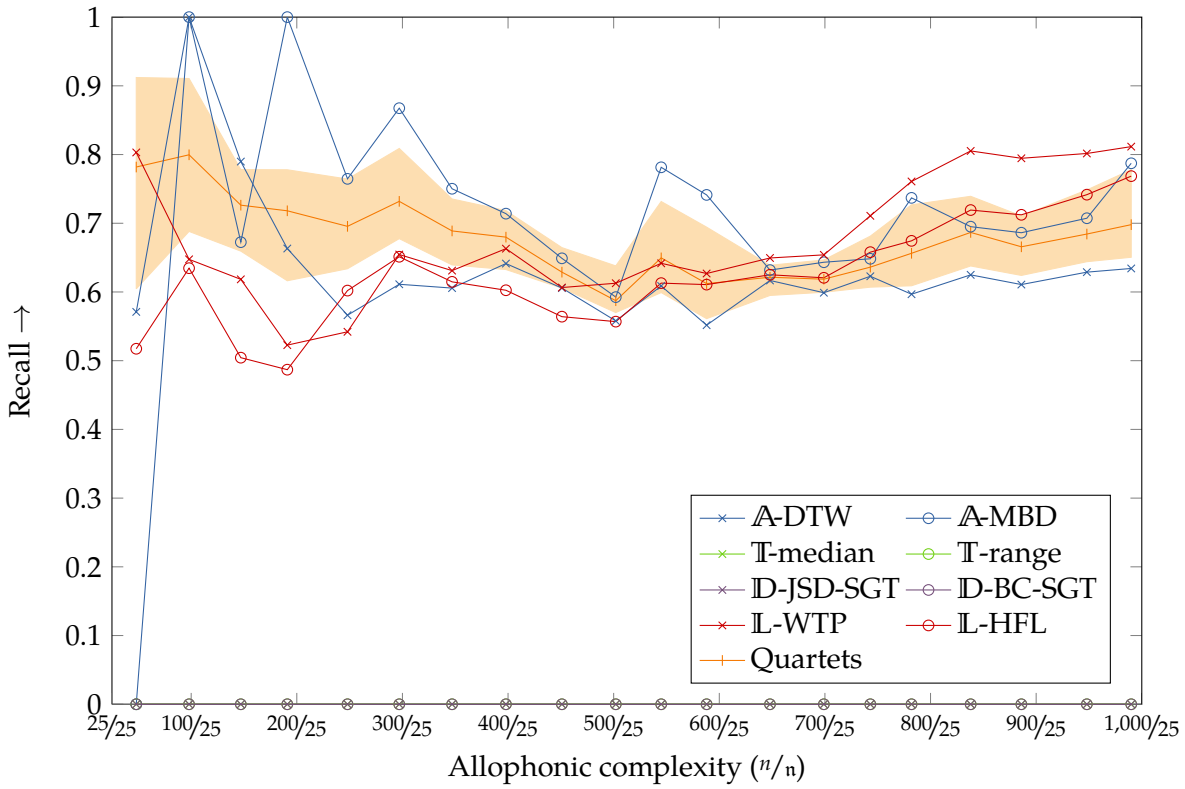


Figure 4.24 — Recall of the logistic regression models on the binary classification task. For quartet models, the solid line marks the average recall, and the band is bounded by the lowest and the highest values for each allophonic complexity.

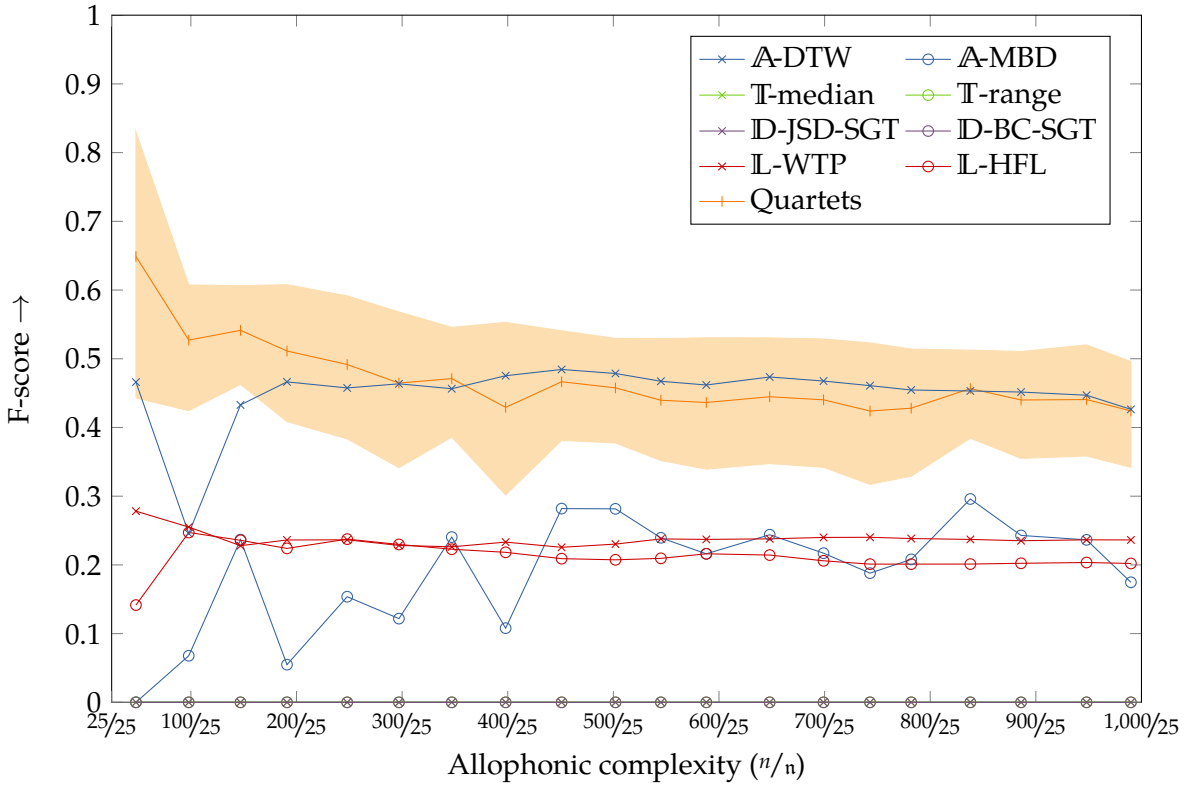


Figure 4.25 — F-score of the logistic regression models on the binary classification task. For quartet models, the solid line marks the average F-score, and the band is bounded by the lowest and the highest values for each allophonic complexity.

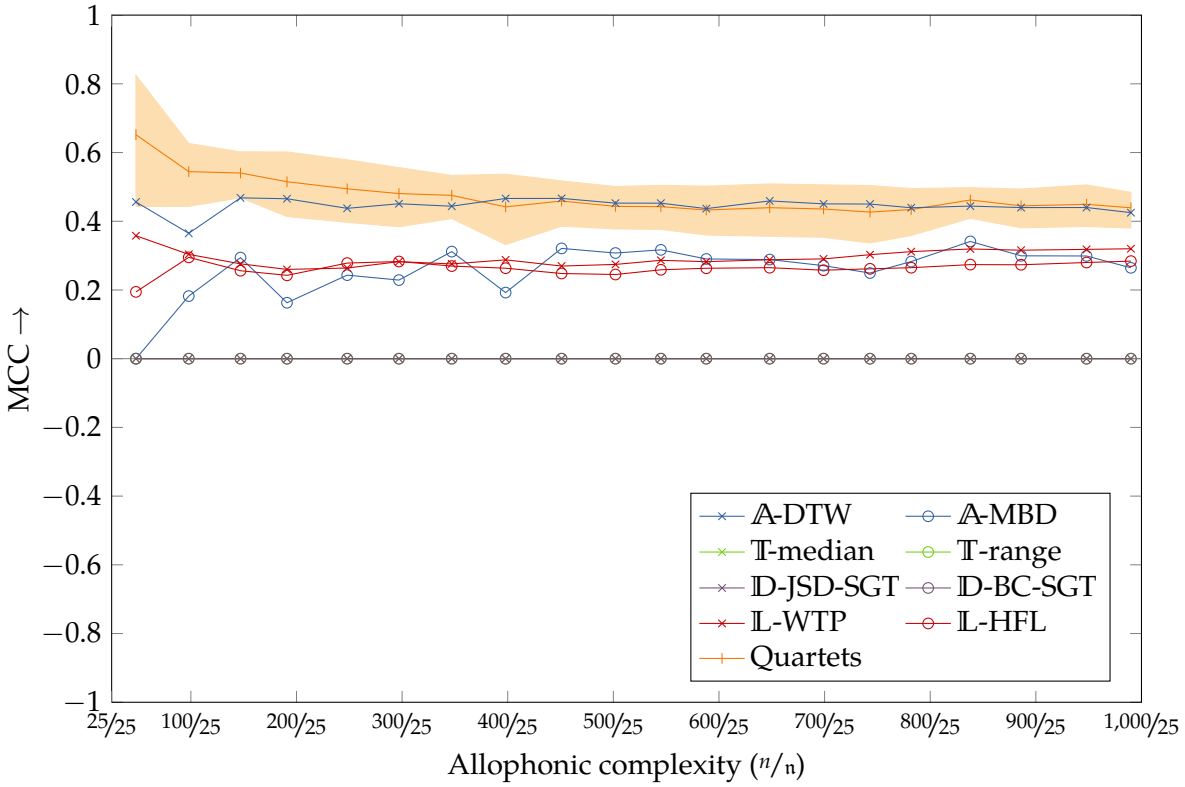


Figure 4.26 — MCC of the logistic regression models on the binary classification task. For quartet models, the solid line marks the average MCC, and the band is bounded by the lowest and the highest values for each allophonic complexity.

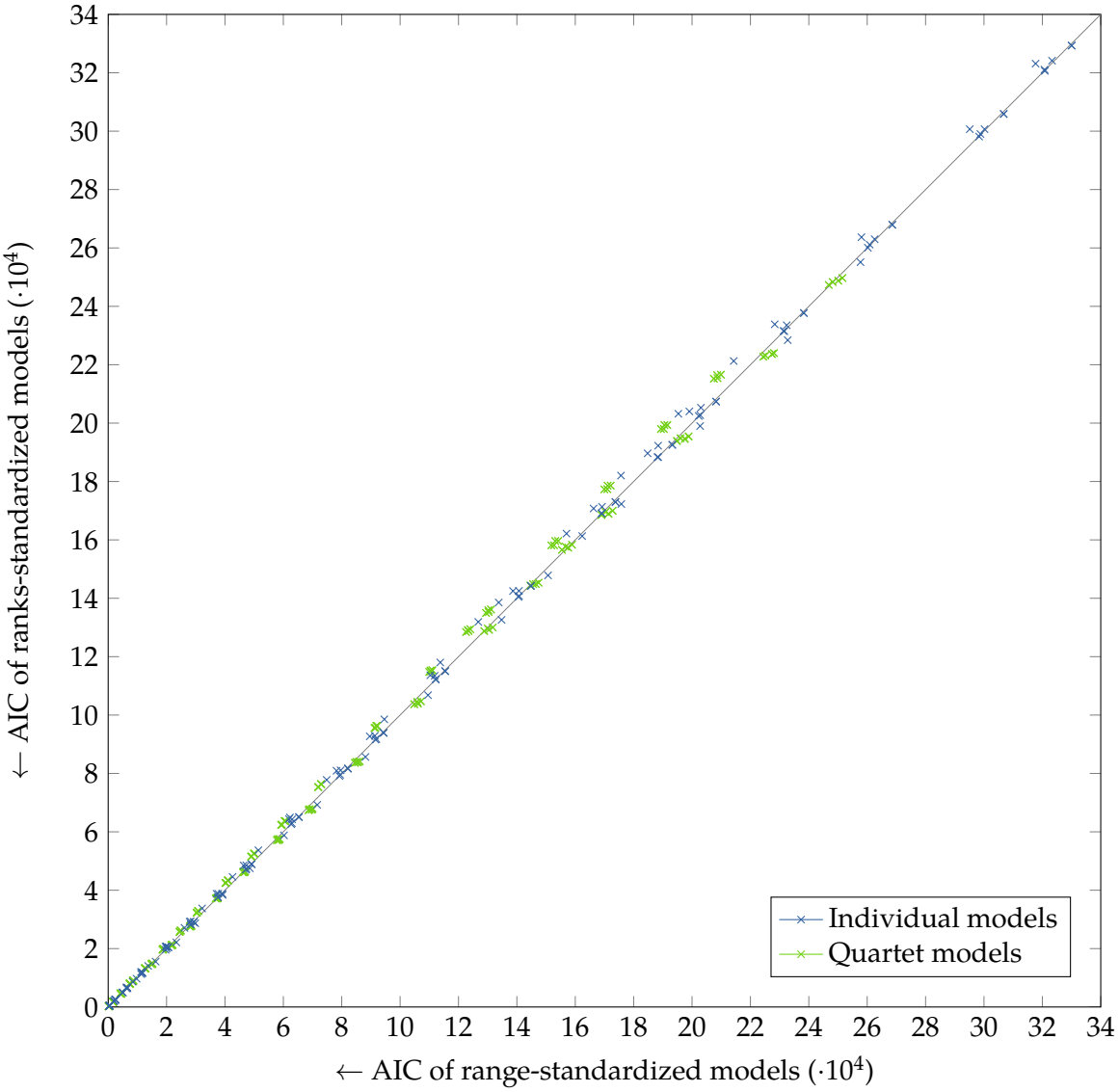


Figure 4.27 — Relative influence of range- and ranks-based standardization on the goodness of fit of the binomial logistic regression models of allophony, across all allophonic complexities. The gray line denotes the identity function.

TN. Therefore, the performance of temporal and distributional indicators in terms of accuracy, as reported in Figure 4.22, can be interpreted as the baseline performance for any model of the acquisition of allophony. Indeed, observing accuracy values below those of 𝕋-median, 𝕋-range, 𝔻-JSD-SGT, and 𝔻-BC-SGT would indicate that not only were all allophonic pairs classified as non-allophonic, but also that some non-allophonic pairs were classified as allophonic. As reported in Figure 4.26, null MCC values confirm that temporal and distributional indicators yield predictions that are not better correlated with the true allophonic statuses than random predictions: all four indicators lie on the chance line, regardless of the allophonic complexity.

Furthermore, it is worth noting that the lexical indicators 𝕃-WTP and 𝕃-HFL appear to be almost as effective as the acoustic indicator 𝔸-MBD, especially with regards to the F-score and MCC values presented in Figures 4.25 and 4.26. Whereas they do not compete with 𝔸-DTW or any quartet models, lexical indicators seem to be much more effective indicators of allophony than what was suggested by their ρ prognosis or the examination of their (extremely skewed) distributions in Section 4.4. As previously stated, using lexical indicators in models of early language acquisition requires accounting for the acquisition of the lexicon; yet, these classification

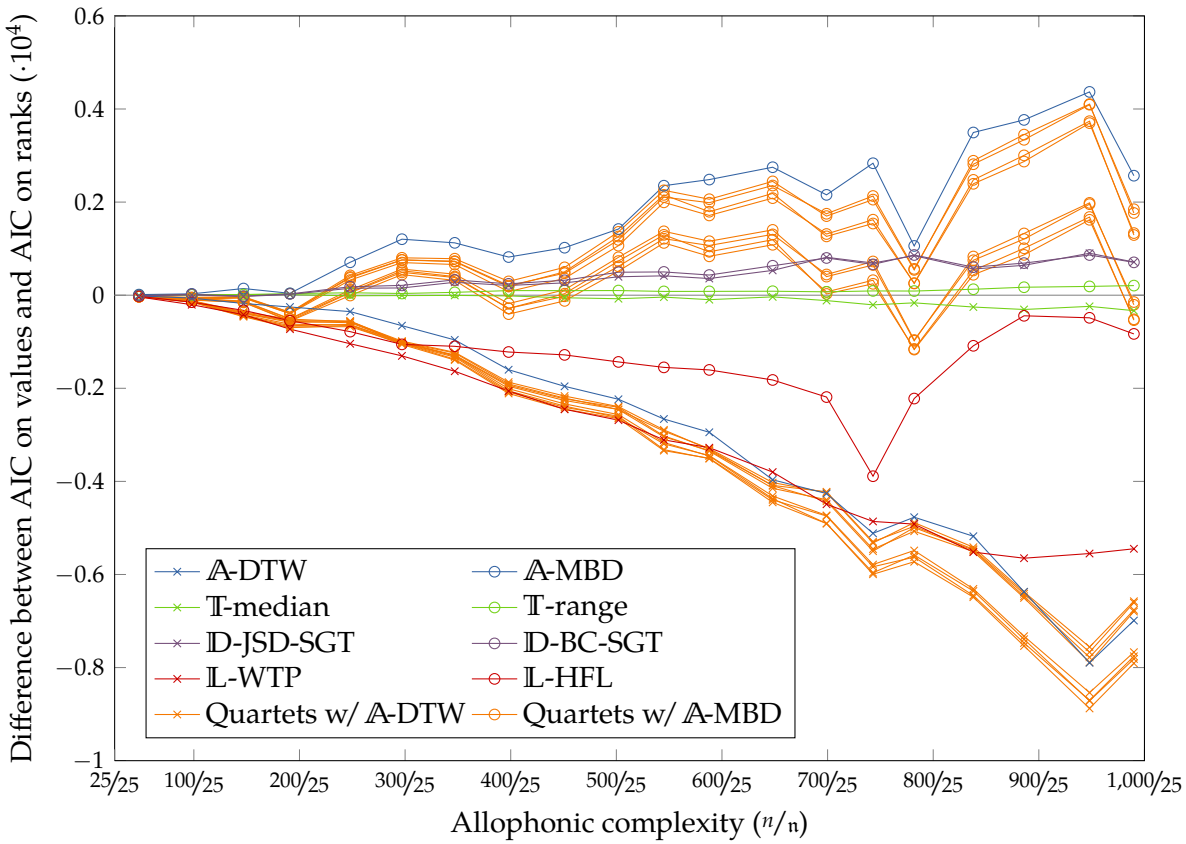


Figure 4.28 — Relative influence of range- and ranks-based standardization on the goodness of fit of the binomial logistic regression models of allophony, as a function of allophonic complexity. The gray line denotes the identity function.

results strengthen Martin et al.'s (2009) proposal that using top-down information might prove to be relevant for the discovery of allophony.

Range or ranks? As previously mentioned, about half of the studies building upon Peperkamp et al.'s (2006) original experiments limited the evaluation of their (yet to be named) indicators of allophony to the rank-based prognosis of allophony we refer to as ρ (Martin et al., 2009; Dautriche, 2009; Boruta, 2009). Nonetheless, indicators of allophony do not only provide a ranking of all possible phones pairs in a given allophonic inventory, they also provide phone-to-phone dissimilarity ratings. Furthermore, as we observed in Section 4.4.3, all indicators of allophony have non-uniform distributions. We are thus interested in assessing whether allophony is better learned from a mere ranking of an indicator's values or from an examination of these non-uniformly distributed values. In order to do so, all aforementioned binomial logistic regression models were duplicated and trained using either range- or rank-standardized indicators.

In order to compare both setups, we examine the AIC values of matching binomial logistic regression models. As presented in the scatter plot Figure 4.27 for all models and allophonic complexities, there is no significant difference between both standardization techniques in terms of the goodness of fit of the models. Indeed, all points in that plot lie exactly on or close to the diagonal line representing the identity function.

However, as illustrated in Figure 4.28, patterns emerge when comparing the goodness of fit of range- and ranks-standardized models of allophony as a function of the allophonic complexity of their input. In this figure, the values we plotted against the allophonic complexity are the differences between the AIC of range-standardized models and that of ranks-standardized models. Because AIC is to be minimized, positive differences indicate that the AIC observed

on values is lesser than the AIC observed on ranks and, hence, that using the true values of the corresponding indicator(s) yield a better-fitting model. Surprisingly, temporal indicators are insensitive to both allophonic complexity and the standardization of their input. This observation is however supported by the confusion plots in Section 4.4.3 and, especially, Figure 4.16 whereby 𝕋-median and 𝕋-range appear to have the most uniform distributions (yet shapeless). The most striking pattern in Figure 4.28 is, however, that the goodness of fit values for quartet models follow two strong, distinct trends depending on the acoustic indicator they comprise. On one hand, 𝔸-DTW and quartets that include 𝔸-DTW seem to benefit from a mere ranking of the values. On the other hand, 𝔸-MBD and quartets that include 𝔸-MBD seem to benefit from the true values of the indicators. Yet, both acoustic indicators have comparable confusion plots, as illustrated in Figure 4.15. Further research is thus needed to take a stance on the issue of standardization, as will be discussed in Chapter 5.

4.6 Overall assessment

In this chapter, we defined a common framework for all phone-to-phone (dis)similarity measures proposed in studies building upon Peperkamp et al.'s original experiments (Peperkamp et al., 2006; Le Calvez, 2007; Le Calvez et al., 2007; Dautriche, 2009; Martin et al., 2009; Boruta, 2009, 2011b; this study). We then carried out an extensive quantitative evaluation of the performance of various indicators relying on acoustic, temporal, distributional, or lexical information. For instance, the experiments reported in Section 4.5 give, to our knowledge, the first empirical bounds on the learnability of the allophony relation in this framework.

In these experiments, we showed that whereas classifying all possible phone pairs in a given allophonic inventory appears to be quite an easy task, the high performance we observed is mostly due to the correct rejection of non-allophonic pairs. Moreover, as we argued throughout this chapter, non-allophonic pairs are simply the spurious byproduct of a combinatioral enumeration as, in the perspective of the discovery of phonemes, only allophonic pairs are informative. Furthermore, it is worth noting that solving the task introduced by Peperkamp et al. (2006) does not provide a complete answer to the problem of the discovery of the phonemic inventory of a given language as, in fact, an additional step would be required to reconstruct each and every phoneme from the structureless set of all allophonic pairs.

For these reasons, the experiments to be presented in the next chapters depart from Peperkamp et al.'s original experiments as we will specifically adress the problem of learning the polyadic relation of phonemehood—that is to say, the whole phonemic inventory of the target language. In other words, do indicators of allophony contain information about phonemehood?

CHAPTER 5
INDICATORS OF PHONEMEHOOD

The goal of this chapter is to assess whether previously defined indicators of allophony are also indicators of phonemehood. Concretely, we are interested in evaluating the informativeness of these indicators when the task does not only consist in predicting whether two phones are allophones but, also, determining of which phoneme they both are realizations. Indeed, whether it is from the point of view of theoretical, computational, or developmental linguistics, learning the allophony relation is only the byproduct of or—following Peperkamp et al. (2006)—a preliminary step to the discovery of the phonemic inventory of the target language. However, for the sake of comparability with previously reported experiments building upon Peperkamp et al.'s pairwise framework—including the ones we presented in Chapter 4—we take an intermediate step in this chapter, as we aim at predicting phoneme-like categories while retaining the pairwise formulation of the indicators.

This chapter is divided into three main sections. In Section 5.1, we discuss the formal and empirical limitations of the pairwise framework, as well as the concepts and data structures we use throughout this chapter. In Section 5.2, we then report various classification experiments aimed at providing empirical upper bounds on the learnability of phonemehood in a pairwise framework. Finally, Section 5.3 contains a general assessment of the work presented in this transitional chapter.

5.1 Phonemehood: definitions and objectives

As discussed in Chapters 2 and 4, the phonemic inventory of a given language comprises a finite number of abstract sound categories (i.e. phonemes) whose tangent realizations in a word or utterance are conditioned by the surrounding phonemes (i.e. allophones). Following the work of Peperkamp et al. (2006), and as presented in Chapter 3, we adopted a quantized representation of the input data whereby the variability in phoneme realizations is reduced to a finite set of phones, a.k.a. the allophonic inventory. Because we further assumed that no two phonemes may have common realizations (cf. Assumption 3.3), the phonemic inventory $\mathfrak{P}$ of the target language can be formally defined as a partition of the allophonic inventory P. In other words, $\mathfrak{P}$ is a set of non-empty subsets (i.e. the phonemes) of P such that every phone $p_i \in P$ belongs to exactly one of these subsets.

As a preliminary step to the discovery of the phonemic inventory $\mathfrak{P}$, Peperkamp et al. (2006) introduced a pairwise task whereby the learner has to decide—for each of the $n(n-1)/2$ possible pairs of phones in P—whether or not two phones are allophones, hence defining a binary relation over P that we refer to as allophony. However, no linguistic theory would define phonemes as having exactly two realizations; as we observed in Chapter 3, the speech-based procedure we used for the computation of putative allophonic inventories yielded phonemes with varied number of allophones. In this chapter, we are thus interested in shifting the focus from *allophony*

to *phonemehood*, i.e. our goal is to learn phoneme-like categories resembling the true phonemic partition $\mathfrak{P}$.

5.1.1 Limitations of the pairwise framework

In Chapter 4, we built upon Peperkamp et al.'s (2006) pairwise framework using—for the sake of comparability—the same assumptions, task, and (to a certain extent) methods without discussion. However, when shifting the learning objective from allophony to phonemehood, some empirical and formal limitations arise. In this section, we examine the major limitations of the pairwise framework.

The problem of transitivity We previously assumed that phone pairs are independent from one another, meaning that the allophonic status $a^*_{ij} \in \mathbb{B}$ (i.e. allophonic or non-allophonic) of a given phone pair $\{p_i, p_j\} \subseteq P$ makes it neither more nor less probable that another one is allophonic or not (cf. Assumption 4.2). In fact, all aforementioned studies building upon Peperkamp et al.'s (2006) framework made this—implicit—assumption (Le Calvez, 2007; Le Calvez et al., 2007; Dautriche, 2009; Martin et al., 2009; Boruta, 2009, 2011b). This assumption is, however, far from being trivial, as the two possible allophonic statuses do not actually describe a binary partition of the allophonic inventory P, but a binary relation over P. Indeed, whereas previous experiments treated phone pairs such as $\{p_i, p_j\} \subseteq P$ or $\{p_i, p_{j'}\} \subseteq P$ as atomic objects, a rigorous model of allophony should not ignore that they both include the phone p_i.

Furthermore, let us recall a more fundamental assumption: no two phonemes have common realizations (cf. Assumption 3.3). Therefore, the symmetric binary relation over P that we refer to as *allophony* actually describes a (hopefully $\mathfrak{n}$-ary) partition of P into non-empty and mutually-exclusive putative phonemes, and the relation that we refer to as *phonemehood* describes the property shared by some phones of belonging to a given phoneme-like subset of this partition. Because we assumed that phonemehood describes a crisp partition of P, allophony can be but a transitive relation, that is

$$(a_{ij} = 1) \wedge (a_{ij'} = 1) \Rightarrow (a_{jj'} = 1). \tag{5.1}$$

However, the assumption of independence and this (indirect) assumption of transitivity are contradictory, as the former negates the underlying structure in the data that the latter demands.

This problem arises from Peperkamp et al.'s (2006) formulation of the task whereby the first step in the discovery of phonemes is the discovery of allophonic pairs. Even though learning which pairs of phones are realizations of the same phoneme is not irrelevant for the task at hand, it does not give a complete answer to the problem of learning phonemic categories. Indeed, whereas phonemehood is a polyadic relation that exhaustively describes the realizations of every phoneme in the target language, allophony is a binary relation that merely indicates whether two phones are realizations of the same phoneme but—in that event—contains no information regarding the identity of the phoneme of which they both are realizations. Learning allophony is thus the first step of a more complex procedure. In that case, however, the formal problem of transitivity arises. By contrast, directly learning phonemehood from the dissimilarity representations given by the indicators would not yield such a problem, as it would involve learning a partition of a set—rather than deriving a partition over a set from a binary relation over that set. In this chapter and the next, we precisely aim at learning phonemehood directly from our indicators' data.

It is worth mentioning that one answer to the problem of transitivity would consist in computing the transitive closure of the binary relation described in a given matrix of allophony $\mathbf{A}$, for example using the Floyd–Warshall algorithm (Floyd, 1962; Warshall, 1962; Cormen et al., 2001). However, we chose not to explore such an approach to phonemehood in the present study, on the grounds that transitive closure would propagate false positives in $\mathbf{A}$—i.e. non-allophonic pairs wrongfully considered to be allophonic—and may eventually describe a binary relation over P whereby all phones are allophones of a unique phoneme.

Table 5.1 — Hapax phonemena: for each allophonic complexity, the listed phonemes have only one realization—with the consequence that they only appear in non-allophonic pairs. In the last column, the hapax rate is given by the percentage of phonemes in the inventory that have only one realization.

Complexity	Hapax phenomena	Hapax rate
$^{n}/_{\mathfrak{n}} \leq {}^{48}/_{25}$	{u, aː, eː, iː, uː, w, j, b, d, g, p, m, ɴ, s, z, r, h}	68%
$^{n}/_{\mathfrak{n}} \leq {}^{98}/_{25}$	{aː, iː, uː, w, b, d, g, p, ɴ, z, r, h}	48%
$^{n}/_{\mathfrak{n}} \leq {}^{147}/_{25}$	{aː, iː, w, b, d, g, p, z, r, h}	40%
$^{n}/_{\mathfrak{n}} \leq {}^{191}/_{25}$	{aː, iː, w, b, p, z, r, h}	36%
$^{n}/_{\mathfrak{n}} \leq {}^{248}/_{25}$	{iː, w, b, p, z, h}	24%
$^{n}/_{\mathfrak{n}} \leq {}^{502}/_{25}$	{iː, w, b, p}	16%
$^{n}/_{\mathfrak{n}} \leq {}^{948}/_{25}$	{iː, p}	8%
$^{n}/_{\mathfrak{n}} \leq {}^{990}/_{25}$	{iː}	4%

The problem of hapax phenomena Predicting of which phoneme both phones in a given allophonic pair are realizations presents another problem, yet purely empirical. If a given phoneme has no more than one realization in the allophonic inventory P, then this phoneme's phones are only involved in non-allophonic pairs and, consequently, the phoneme never occurs as the category to which the two phones of a pair belong.

As presented in Table 3.7 and further illustrated in Table 5.1, this situation is attested in our datasets for 18 of the 25 phonemes of Japanese and—even at the highest allophonic complexity considered in this study—the problem of hapax phonenomena (i.e. unique events) still occurs for the vowel /iː/. On one hand, this empirical problem could be seen as a consequence of an over-quantized input (cf. Assumption 3.1) where—part of—the problem of discovering phonemes has already been solved, as some phonemes are presented in the input at their highest possible degree of abstraction. On the other hand, allophones were computed in Section 3.2.2 using a procedure that quantizes at once all the realizations of all phonemes in the CSJ so that, for any desired allophonic complexity, the output allophonic inventory is given by the quantized partition of the data that best accounts for the acoustic variability in the speech recordings. Hence, as we argued in Section 3.3.2, if a given phoneme was only assigned one realization while others were assigned dozens, we can fairly assume that this phoneme's realizations show very little (acoustic) variability. Thus, the only way for a pairwise framework to observe an allophonic pair made up of two realizations of such a low-variability phoneme would be to use non-quantized input—i.e. raw acoustic signal—but the problem would then become computationally intractable for significantly large datasets, notwithstanding the consequent impossibility of relying on distributional or lexical indicators, as discussed in Section 4.2.3.

5.1.2 Objectives: predicting phonemehood

As we will explain in Chapter 6, learning the phonemic partition $\mathfrak{P}$ of a given allophonic inventory P requires a complete reformulation of the learning framework. In this chapter—for the sake of comparability with previously reported experiments—we take an intermediate step, and assess the performance of our indicators of allophony in a tentative multiclass extension of Peperkamp et al.'s (2006) framework.

The binary task of allophony introduced by Peperkamp et al. consists in predicting, for every possible pair of phones in a given allophonic inventory, whether or not two phones are allophones (cf. Figure 4.19). Here, we aim at predicting, for every possible pair of phones in the allophonic inventory, whether or not two phones are allophones and, in that case, of which phoneme they both are realizations. There are thus $\mathfrak{n}+1$ different target classes in the latter task: the $\mathfrak{n}$ phonemes $\mathfrak{P} \equiv \{\mathfrak{p}_1, \ldots, \mathfrak{p}_\mathfrak{n}\}$ of the target language as well as an additional dummy class, denoted $\otimes$, used as the phoneme-level class of non-allophonic pairs.

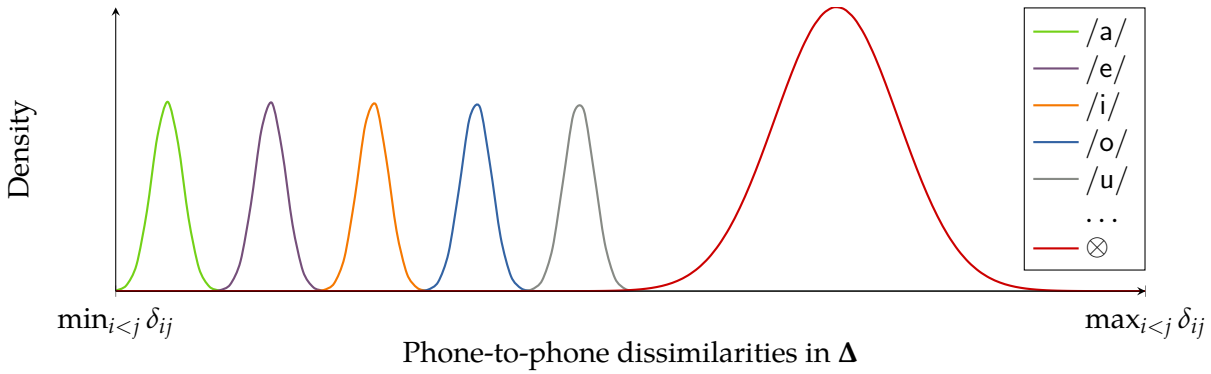


Figure 5.1 — Schematic representation of the distribution of an ideal indicator of phonemehood.

A pairwise framework Amending the task introduced by Peperkamp et al., our goal is now to learn a $n \times n$ symmetric matrix of phonemehood $\mathbf{\Lambda} \equiv [\lambda_{ij}]$ where $\lambda_{ij} \in \mathfrak{P} \cup \{\otimes\}$ denotes the phoneme-like category of a given pair of phones $\{p_i, p_j\} \subseteq P$. Let $\mathbf{\Lambda}^* \equiv [\lambda^*_{ij}]$ denote the reference matrix of phonemehood where λ^*_{ij} is the reference phoneme-like category of the pair of phones $\{p_i, p_j\} \subseteq P$, given by

$$\lambda^*_{ij} \equiv \begin{cases} \mathfrak{p}_h & \text{if } \exists\ \mathfrak{p}_h \in \mathfrak{P}, \{p_i, p_j\} \subseteq \mathfrak{p}_h \\ \otimes & \text{otherwise} \end{cases} . \tag{5.2}$$

The reference matrix of allophony $\mathbf{A}^*$ and the reference matrix of phonemehood $\mathbf{\Lambda}^*$ are similar in the sense that all values equal to 1 in $\mathbf{A}^*$ (i.e. denoting allophonic pairs) are further specified by the appropriate phonemes in $\mathbf{\Lambda}^*$ and, mutatis mutandis, all values equal to 0 in $\mathbf{A}^*$ (i.e. denoting non-allophonic pairs) are merely replaced by $\otimes$ in $\mathbf{\Lambda}^*$.

Indicators of phonemehood In order to predict the phoneme-like class of any pair of phones in a given allophonic inventory, we rely on (combinations of) the indicators of allophony we introduced and benchmarked in Chapter 4, viz. 𝔸-DTW, 𝔸-MBD, 𝕋-median, 𝕋-range, 𝔻-JSD-SGT, 𝔻-BC-SGT, 𝕃-WTP, and 𝕃-HFL. Although all indicators were, by definition, computed to be—to the greatest extent possible—correlated with allophony, none of them were purposedly designed so that specific subranges of its values would be further correlated with specific phonemes, as illustrated in Figure 5.1. Moreover, it is worth highlighting that having access to an ideal matrix of allophony $\mathbf{A} = \mathbf{A}^*$ is not sufficient to retrieve the phonemic inventory of the target language. Returning to our prior analogy between the allophony relation and the siblings relationship (cf. Section 4.1), should given siblings share genes, eye color, or blood type, such features would be insufficient to identify their parents. In other words, knowing that two individuals are siblings does not indicate to which sibship they belong. Be that as it may, we test in this chapter the hypothesis that indicators of allophony also contain information regarding phonemehood.

Finally, because the results we presented in Section 4.5.3 were inconclusive on the issue of standardization, we consider both range- and ranks-standardized indicators in this chapter.

Prognosis of phonemehood In our multiclass, $(\mathfrak{n}+1)$-ary extension of Peperkamp et al.'s (2006) framework, the output data comprises both an indirect, putative partition of the allophonic inventory (i.e. the phoneme-specified allophonic pairs) and additional information regarding all pairs of phones classified as having no relevant relation for the definition of phonemes (i.e. the $\otimes$-classified non-allophonic pairs). Unfortunately, to our knowledge, no statistic has been proposed so far that would yield a prognosis of phonemehood for such data. Indeed, as will be further discussed in Section 6.2, so called cluster validity indices (Bezdek & Pal, 1998; Kim & Ramakrishna, 2005; Aliguliyev, 2009; among others) rely on a true partition of the input data.

Nonetheless, it is worth mentioning that the probabilistic prognoses of binary allophony we presented and discussed in Section 4.4.1 bear some relevance to our now-multiclass pairwise

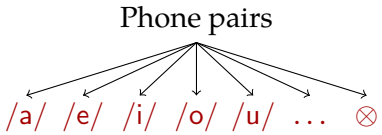


Figure 5.2 — Task diagram: flat-response $(\mathfrak{n}+1)$-ary phonemehood on phone pairs. The task consists in predicting, for each possible phone pair in a given allophonic inventory, of which phoneme-like category both phones are realizations—i.e. the underlying phoneme for allophonic pairs, or the dummy category $\otimes$ for non-allophonic pairs.

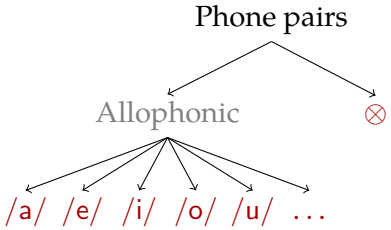


Figure 5.3 — Task diagram: nested-response $(\mathfrak{n}+1)$-ary phonemehood on phone pairs. The task consists in predicting whether or not two phones are realizations of the same phoneme and, in that case, of which phoneme they both are realizations.

problem. Quantitatively speaking, the prominent class in both tasks is the class comprising all non-allophonic pairs—as illustrated in Figure 4.5 and the confusion plots in Section 4.4.3. Moreover, Herrnstein et al.'s (1976) ρ is an estimate of the probability of a randomly-drawn non-allophonic pair being more dissimilar than a randomly-drawn allophonic pair in the dissimilarity ratings of a given indicator of allophony. This rank-sum test therefore assesses the separation of allophonic and non-allophonic pairs and, as a consequence, provides a partial quantitative assessment of the separability of non-allophonic pairs from the $\mathfrak{n}$ phoneme-like classes. The separability of the remaining $\mathfrak{n}$ phoneme-like classes is, however, unaccounted for by this statistic. For these reasons, we can speculate that the most effective indicators of allophony (viz. acoustic indicators) are also the most effective indicators of phonemehood.

5.2 Predicting phonemehood: $(\mathfrak{n}+1)$-ary classification task

The classification experiments we report in this section are straightforward multiclass extensions of the ones we reported in Section 4.5. More precisely, for each possible phone pair in a given allophonic inventory, we not only aim to predict whether or not both phones are allophones but also, in that case, to predict of which phoneme they both are realizations.

To this aim, we compare the performance of two types of multiclass regression models. Indeed, the $\mathfrak{n}+1$ phoneme-like classes we defined in the previous section can either be considered as the comparable outcomes of a flat, single-level multinomial variable (as illustrated in Figure 5.2), or as the incomparable outcomes of a nested, two-level multinomial variable (as illustrated in Figure 5.3). In the latter case, the model is to be interpreted as a refinement of the binomial model we presented in Section 4.5, as predicting of which phoneme two given phones are realizations is only meaningful if these phones form an allophonic pair. However, the performance of this nested-response model is thus necessarily bounded by the performance of the top-level binomial model. If a given allophonic pair is first misclassified as being non-allophonic, the predicted class is inevitably $\otimes$. Conversely, if a given non-allophonic pair is first misclassified as being allophonic, the model will then inevitably assign a phoneme-like category instead of $\otimes$.

Although it merely negates the inherently nested nature of pairwise phonemehood, we also consider the simpler, flat-response model as its performance is not bounded by that of the binomial model we presented in Section 4.5. In the context of logistic regression models, the

flat-response model is a classic and straightforward multinomial extension of the binomial logistic regression model. By contrast, the literature on nested-response multinomial logistic regression is, to our knowledge, scarce—as we will discuss in Section 5.2.2.

5.2.1 Flat-response multinomial logistic regression

As presented in Section 4.5.1, binomial logistic regression can be used to model a binary response variable based on an arbitrary number of predictor variables. When the number of distinct outcomes exceeds two, generalizations of the logistic regression model to nominal (i.e. unordered) multinomial response variables are readily available (Faraway, 2006; Agresti, 2007). Using the same definition and notation for the input data as we used in the binomial case, let $\{\pi_{ij(0)}, \pi_{ij(1)}, \ldots, \pi_{ij(\mathfrak{n})}\}$ denote the response probabilities associated to the $\mathfrak{n}+1$ phoneme-like classes $\mathfrak{C} \equiv \{\otimes, \mathfrak{p}_1, \ldots, \mathfrak{p}_\mathfrak{n}\}$ for a given phone pair $\{p_i, p_j\} \subseteq P$, where

$$\pi_{ij(0)} \equiv \mathrm{P}(\lambda^*_{ij} = \otimes \mid \delta_{ij1}, \ldots, \delta_{ij\kappa}) \tag{5.3}$$

denotes the probability of both phones not being realizations of the same phoneme given the κ dissimilarity ratings $\delta_{ij1}, \ldots, \delta_{ij\kappa}$, and

$$\pi_{ij(h)} \equiv \mathrm{P}(\lambda^*_{ij} = \mathfrak{p}_h \mid \delta_{ij1}, \ldots, \delta_{ij\kappa}) \tag{5.4}$$

denotes the probability of both phones p_i and p_j being realizations of the phoneme $\mathfrak{p}_h$ given the same dissimilarities. The $\mathfrak{n}+1$ response probabilities satisfy

$$\pi_{ij(0)} + \sum_h \pi_{ij(h)} = 1. \tag{5.5}$$

Multinomial logistic regression models rely on all pairs of outcome categories and specify the odds of outcome in one category instead of another (Agresti, 2007; pp. 173–179). Concretely, a multinomial logistic regression model pairs each category with a baseline category. Let the dummy category $\otimes$ be the baseline category in the following definitions; it is worth noting, however, that the practical choice of the baseline category has no consequence for the estimation of the model's parameters—*baseline* does not entail *dummy*, and vice versa. This being said, the baseline-category logits are defined as

$$\log\left(\frac{\pi_{ij(h)}}{\pi_{ij(0)}}\right) \equiv \alpha_h + \sum_k \beta_{hk}\delta_{ijk} \tag{5.6}$$

where β_{hk} is the regression parameter linking the phoneme $\mathfrak{p}_h$ to the k-th indicator. Given that the response falls in category $\mathfrak{p}_h$ or $\otimes$, the quantity in Equation 5.6 is to be interpreted as the log-odds that the response is $\mathfrak{p}_h$. The whole mutinomial model has $\mathfrak{n}$ such equations—with separate parameters for each—that are fitted simultaneously. It is worth mentioning that the multinomial logistic regression model has an alternative expression in terms of the response probabilities (Agresti, 2007), given by

$$\pi_{ij(h)} \equiv \frac{\exp(\alpha_h + \sum_k \beta_{hk}\delta_{ijk})}{\exp(\alpha_0 + \sum_k \beta_{0k}\delta_{ijk}) + \sum_{h'} \exp(\alpha_{h'} + \sum_k \beta_{h'k}\delta_{ijk})} \tag{5.7}$$

where the denominator is the same for each category, and the numerators for all possible $\mathfrak{n}+1$ categories sum to the denominator, thus satisfying Equation 5.5.

Multinomial logistic regression models are strictly equivalent to so called maximum entropy models (a.k.a. MaxEnt models; Klein & Manning, 2003; Jurafksy & Martin, 2009; Mount, 2011). Although the latter name is more common in the fields of computational linguistics and natural language processing, we use the former to emphasize the relationship between such models and the binomial logistic regression models we used in Chapter 4.

Underlying assumptions The assumptions of the binomial and the multinomial logistic regression models regarding the response and predictor variables are virtually identical:

— no important predictor variables are omitted, and no extraneous predictor variables are included (cf. Assumption 4.4);

— the predictor variables are not linear combinations of each other (cf. Assumption 4.5);

— the predictor variables are measured without error (cf. Assumption 4.6).

The last two assumptions, however, need to be amended in order to account for the specific formulation of the multinomial model:

Assumption 5.1 The true outcome probabilities follow a multinomial distribution. Thereby, the observations are independent, meaning that the outcome for a given observation does not affect the outcome for another (cf. Assumption 4.2).

Assumption 5.2 The log-odds of the true outcomes against the baseline category are a function of a (weighted) linear combination of the predictor variables (cf. Assumption 4.3).

We consider that the arguments we presented in Section 4.5.1 regarding the validity of these assumptions in the binomial case also hold in the multinomial case.

Model-fitting Fitting a multinomial logistic regression model consists in estimating the $\mathfrak{n}\kappa$ regression parameters $\beta_{11}, \ldots, \beta_{\mathfrak{n}\kappa}$ as well as the $\mathfrak{n}$ intercepts $\alpha_1, \ldots, \alpha_{\mathfrak{n}}$. As in the case of binomial logistic regression models, the estimation process relies on a given set of training observations of the form $\langle \delta_{ij1}, \ldots, \delta_{ij\kappa}, w_{ij}, \lambda^*_{ij} \rangle$, i.e. comprising the dissimilarity ratings assigned to a given phone pair $\{p_i, p_j\} \in P$, its weight, and its true phoneme-like category. The numerical algorithm we used in these experiments outputs the $\mathfrak{n}(\kappa + 1)$ values that maximize the likelihood of the parameters given the training data (Agresti, 2007; Venables & Ripley, 2002; cf. `nnet::multinom`).

Finally, for the sake of comparability with the experiments reported in Section 4.5, we reiterate our argument about the relevance of confirmatory overfitting and, accordingly, trained and tested all multinomial logistic regression models using all available data.

Predictions For the purpose of classifying all possible phone pairs in a given allophonic inventory P, the predicted phoneme-like category λ_{ij} of a given pair $\{p_i, p_j\} \subseteq P$ is simply defined as the most probable one, i.e.

$$\lambda_{ij} \equiv \mathfrak{c}_{\hat{\tau}} \qquad \text{where} \qquad \hat{\tau} \equiv \arg\max_{\tau \in (0,1,\ldots,\mathfrak{n})} \pi_{ij(\tau)} \tag{5.8}$$

denotes the index of the most probable phoneme-like category.

5.2.2 Nested-response multinomial logistic regression

In the case of pairwise phonemehood, the phoneme-like outcomes form an intrinsically nested variable: the model first needs to determine whether two given phones are allophones and—in that case, and only in that case—of which phoneme they both are realizations. Whereas there is a considerable literature describing logistic regression models whose predictor variables are structured (a.k.a. hierarchical, random-effects, and mixed-effects models; Agresti, 2007; Gelman & Hill, 2007), logistic regression models whose *response variable* is structured are, to our knowledge, only (briefly) discussed by McCullagh & Nelder (1989; pp. 160–164), Faraway (2006; pp. 113–117), and Rodríguez (2007; ch. 6, pp. 11–15).

Our methodology for such nested-response models follows Faraway's analysis of the dataset collected by Lowe et al. (1971) concerning live births with deformations of the central nervous system. Though completely unrelated, this dataset indeed shares many properties with the data at hand in the present study: live births (pairs of phones) can occur with or without deformations of the central nervous system (can be allophonic or not) and, in that case, various deformations (phonemes) are possible. Moreover, in both datasets, the outcome of lesser interest (no deformation and no allophony) numerically dominates all other possible outcomes.

Following Faraway (2006), we separately develop a binomial model for allophony and a multinomial model for phonemehood. Let

$$\pi_{ij} \equiv \mathrm{P}(\lambda^*_{ij} \neq \otimes \mid \delta_{ij1}, \ldots, \delta_{ij\kappa}) \tag{5.9}$$

denote the conditional probability of the pair of phones $\{p_i, p_j\} \subseteq P$ being allophonic given the κ predictor values $\delta_{ij1}, \ldots, \delta_{ij\kappa}$. As discussed in Section 4.5.1, the top-level binomial logistic regression model has linear form for the logit of the probability of allophony, i.e.

$$\text{logit}(\pi_{ij}) \equiv \alpha + \sum_{k=1}^{\kappa} \beta_k \delta_{ijk}. \tag{5.10}$$

Furthermore, let $\{\pi_{ij(1)}, \ldots, \pi_{ij(\mathfrak{n})}\}$ denote the response probabilities associated to the $\mathfrak{n}$ phonemic classes $\mathfrak{P} \equiv \{\mathfrak{p}_1, \ldots, \mathfrak{p}_\mathfrak{n}\}$ for a given allophonic pair of phones $\{p_i, p_j\} \subseteq P$, where

$$\pi_{ij(h)} \equiv \text{P}(\lambda^*_{ij} = \mathfrak{p}_h \mid \delta_{ij1}, \ldots, \delta_{ij\kappa}, \lambda^*_{ij} \neq \otimes) \tag{5.11}$$

denotes the probability of both phones p_i and p_j being realizations of the phoneme $\mathfrak{p}_h$ given the same dissimilarities. The $\mathfrak{n}+1$ response probabilities satisfy

$$\sum_h \pi_{ij(h)} = 1 \tag{5.12}$$

and, as presented in the previous section, the low-level multinomial logistic regression model as the following expression in terms of the response probabilities:

$$\pi_{ij(h)} \equiv \frac{\exp(\alpha'_h + \sum_k \beta'_{hk} \delta_{ijk})}{\sum_{h'} \exp(\alpha'_{h'} + \sum_k \beta'_{h'k} \delta_{ijk})}. \tag{5.13}$$

Finally, as emphasized by McCullagh & Nelder (1989), it is worth noting that

> ❝There is no good reason here to expect that the coefficients [α, $\beta_1, \ldots, \beta_\kappa$, and $\alpha'_1, \ldots, \alpha'_\mathfrak{n}$, $\beta'_{11}, \ldots, \beta_{\mathfrak{n}\kappa}$] might be equal or even comparable.❞

Predictions For the purpose of classifying all possible phone pairs in a given allophonic inventory P, the predicted phoneme-like category λ_{ij} of a given pair $\{p_i, p_j\} \subseteq P$ is given by

$$\lambda_{ij} \equiv \begin{cases} \otimes & \text{if } \pi_{ij} \leq .5 \\ \mathfrak{p}_{\hat{h}} & \text{otherwise, where } \hat{h} \equiv \arg\max_h \ \pi_{ij(h)} \end{cases} \tag{5.14}$$

denotes the index of the most probable phoneme for the pair of phones $\{p_i, p_j\}$.

5.2.3 Evaluation

In these experiments, as in the ones we presented in Section 4.5, we are interested in the predictive power of the regression models.

Goodness of fit The first quantitative criterion we use to assess the fit of a given multinomial logistic regression model to the reference matrix of phonemehood $\mathbf{\Lambda}^*$ is the aforementioned AIC.

The definition of the AIC in the flat-response multinomial case is a straightforward extension of the definition given in Equation 4.45 for the binomial case: the AIC of a regression model is a measure of the maximized likelihood of its parameters, penalized by the number of such parameters. The only difference lies in the latter penalty: whereas a binomial logistic regression model has $\kappa + 1$ parameters, a multinomial model has $\mathfrak{n}(\kappa + 1)$ parameters.

To our knowledge, no extension of the definition of the AIC has yet been proposed for nested-response multinomial logistic regression models.

Contingency table The other two criteria we use to assess the performance of (combinations of) indicators of phonemehood rely on an actual pairwise classification of their input into the $\mathfrak{n}+1$ phoneme-like categories. These criteria are thus applicable to, and meaningful for, both flat- and nested-response multinomial models. For each regression model, a contingency table is used to cross-classify the reference outcomes with the model's predictions. In this multiclass setup, this cross-classification yields a $(\mathfrak{n}+1) \times (\mathfrak{n}+1)$ contingency table. Let $\mathbf{T} \equiv [t_{\tau'\tau}]$ denote such a contingency table where, by convention, τ denotes the index over reference classes and τ' the index over predicted classes. The table is thus non-symmetric: while t_{0h} denotes the number of non-allophonic pairs that were misclassified as realizations of $\mathfrak{p}_h$, t_{h0} denotes the number of

realizations of $\mathfrak{p}_h$ that were misclassified as non-allophonic pairs. In each cell, the non-negative count $t_{\tau'\tau}$ is given by

$$t_{\tau'\tau} \equiv \sum_{i<j} w_{ij} [\![\lambda^*_{ij} = \mathfrak{c}_{\tau'}]\!] [\![\lambda_{ij} = \mathfrak{c}_\tau]\!]. \tag{5.15}$$

Pairwise evaluation The evaluation measures we defined in Section 4.5.2—viz. precision, recall, F-score, and MCC—are only appropriate when there are exactly two possible outcomes. Multiclass predictions can however be considered as a series of pairwise decisions, one for each pair of observations in the test data (Manning et al., 2008). In that case, *TP* denotes the (weighted) number of pairs of observations of equal class that were classified as belonging to the same class, *TN* the number of pairs of observations of different classes that were classified as belonging to different classes, *FP* the number of pairs of observations of different classes that were classified as belonging to the same class, and *FN* the number of observations of equal class that were classified as belonging to different classes.

Because our elementary observations consist in pairs of phones, such a pairwise evaluation actually involves comparing the predictions for all pairs of pairs of phones. For a given allophonic inventory and a given regression model, computing the 2×2 pairwise contingency table thus requires $\mathcal{O}(n^4)$ operations and, hence, becomes computationally intractable as the allophonic complexity of the input increases—especially considering the fact that this computation needs to be repeated for 24 different regression models, and 20 allophonic inventories.

Global accuracy Various evaluation measures can however be computed from a $(\mathfrak{n}+1) \times (\mathfrak{n}+1)$ contingency table. The first measure on which we rely is the global accuracy of the predictions, that is to say the average number of correct predictions across all phoneme-like classes. Formally, global accuracy is given by

$$\text{accuracy}(\boldsymbol{\Lambda}, \boldsymbol{\Lambda}^*) \equiv \mathrm{P}(\lambda_{ij} = \lambda^*_{ij}) \approxeq \frac{\sum_\tau t_{\tau\tau}}{\sum_\tau \sum_{\tau'} t_{\tau'\tau}} = \frac{\sum_{i<j} w_{ij} [\![\lambda_{ij} = \lambda^*_{ij}]\!]}{\sum_{i<j} w_{ij}}. \tag{5.16}$$

In the manner of the pairwise accuracy presented in Section 4.5.2, the global multiclass accuracy can receive a probabilistic interpretation: it is an estimate of the probability $\mathrm{P}(\lambda_{ij} = \lambda^*_{ij})$ of the predicted class λ_{ij} of a randomly-drawn pair of phones $\{p_i, p_j\} \subseteq P$ being equal to its reference phoneme-like class λ^*_{ij}. As far as the contingency table is concerned, a perfectly accurate classification $\boldsymbol{\Lambda} = \boldsymbol{\Lambda}^*$ would result in a strictly diagonal contingency table $\mathbf{T}$. By contrast, the more off-diagonal observations, the less accurate the classification.

Classwise accuracy The second evaluation measure we rely on in these experiments is known as classwise accuracy. For a given phoneme-like class $\mathfrak{c}_\tau \in \mathfrak{C}$, its classwise accuracy is an estimate of the probability $\mathrm{P}(\lambda_{ij} = \mathfrak{c}_\tau \mid \lambda^*_{ij} = \mathfrak{c}_\tau)$ of a randomly-drawn pair of phones $\{p_i, p_j\} \subseteq P$ whose reference class is $\mathfrak{c}_\tau$ being actually classified as $\mathfrak{c}_\tau$. Formally, classwise accuracy is given by

$$\text{accuracy}(\mathfrak{c}_\tau, \boldsymbol{\Lambda}, \boldsymbol{\Lambda}^*) \equiv \mathrm{P}(\lambda_{ij} = \mathfrak{c}_\tau \mid \lambda^*_{ij} = \mathfrak{c}_\tau) \approxeq \frac{t_{\tau\tau}}{\sum_{\tau'} t_{\tau\tau'}} = \frac{\sum_{i<j} w_{ij} [\![\lambda_{ij} = \lambda^*_{ij} = \mathfrak{c}_\tau]\!]}{\sum_{i<j} w_{ij} [\![\lambda^*_{ij} = \mathfrak{c}_\tau]\!]}. \tag{5.17}$$

Whereas global accuracy aggregates all correct predictions—regardless of the predicted class—classwise accuracy allows for a fine-grained exploration of the performance of a multinomial logistic regression model. For instance, we observed in Chapter 4 that the high performance of binomial logistic regression models was mainly due to the correct rejection of non-allophonic pairs. We can thus speculate that the global accuracy of any multinomial model's predictions will benefit from these correct rejections. Comparing the classwise accuracy of $\otimes$ to the classwise accuracies of other phoneme-like categories will thus allow us to assess whether or not indicators of allophony are also indicators of phonemehood.

5.2.4 Results

The performance of all flat-response multinomial logistic regression models is presented in Figure 5.4, Figure 5.6, Figure 5.5 and Table 5.2. The performance of all nested-response multinomial logistic regression models is presented in Figure 5.6 and Table 5.3.

Goodness of fit Let us first consider the goodness of fit tests reported in Figure 5.4 for flat-response models. First, it is worth mentioning that this figure resembles the plot of the AIC of binomial logistic regression models in Figure 4.21. Indeed, as we observed for binomial models, all flat-response models' AIC increases with the allophonic complexity of their input, meaning that the more phonemes have allophones, the less a multinomial logistic regression model is able to fit the phoneme-like partition of all possible pairs of phones. Moreover, acoustic indicators appear to be, once again, the individual indicators that yield regression models with the best fit, as well as the indicators whose AIC gives approximate bounds on that of the quartet models. However, it appears that combining indicators is—on average—no more interesting than only using the acoustic indicator 𝔸-DTW. Finally, individual flat-response models relying on temporal, distributional, and lexical indicators are indistinguishable, as far as AIC is concerned.

However, the goodness of fit tests reported in Figure 5.4 depart from the ones in Figure 4.21 on two main points. First, is it worth noting that the AIC of multinomial models increases faster than that of binomial models. Consider, for example, the AIC values at the maximal allophonic complexity of $^{990}/_{25}$ allophones per phoneme: whereas the AIC of individual binomial models is approximately equal to 35×10^4, the AIC of individual multinomial models is greater than 50×10^4. While, to our knowledge, AIC values are not to be interpreted in isolation, a comparison of both values at $^{990}/_{25}$ allophones per phoneme suffice to show that logistic regression yields models with a fit almost 1.5 times better in the binomial case than in the flat-response multinomial case.

Range or ranks? In the previous chapter, the results presented in Figures 4.27 and 4.28 were inconclusive on the issue of standardization, i.e. assessing whether allophony can be learned from a mere ranking of a given indicator's values, or if the values themselves convey relevant information. To this aim, our methodology has consisted in examining the difference in AIC between pairs of models trained with the same (combination of) indicators, but where one was trained with range-standardized indicators and the other was trained with ranks-standardized indicators. A negative difference indicates that the AIC observed on range-standardized indicators is lesser than the AIC observed on ranks-standardized indicators. Because the AIC is to be minimized, a negative difference indicates that relying on indicators' absolute values yields models with a better fit than relying on mere rankings of each indicator's values or—to say the least—that the $(\mathfrak{n}+1)$-ary phoneme-like classification of all phone pairs is more accurately modeled by a linear combination of values than by a linear combination of ranks.

As illustrated in Figure 5.5, a clear pattern emerges for multinomial logistic regression models from $^{300}/_{25}$ allophones per phoneme onwards. Whereas individual models appear to be rather insensitive to the particular standardization technique used on the indicators (cf. Figure 4.28), we observe significant negative differences between the AIC of range- and ranks-standardized indicators for all quartet models. Moreover, these differences tend to increase with the allophonic complexity of the input. Therefore, phonemehood appears to be better modeled by the true phone-to-phone dissimilarity ratings collected in each indicator of allophony than by the ranking these values denote. This result is important in itself, as the evaluation of various studies building upon Peperkamp et al.'s (2006) framework was limited to the computation of Herrnstein et al.'s (1976) rank-sum test (Martin et al., 2009; Dautriche, 2009; Boruta, 2009, 2011b). Notwithstanding our prior theoretical argument that ρ is at best a prognosis of allophony—rather than a true evaluation measure, cf. Section 4.4.1—we consider the results presented in Figure 5.5 to be empirical evidence against a mere examination of the ranking of phone pairs underlying each

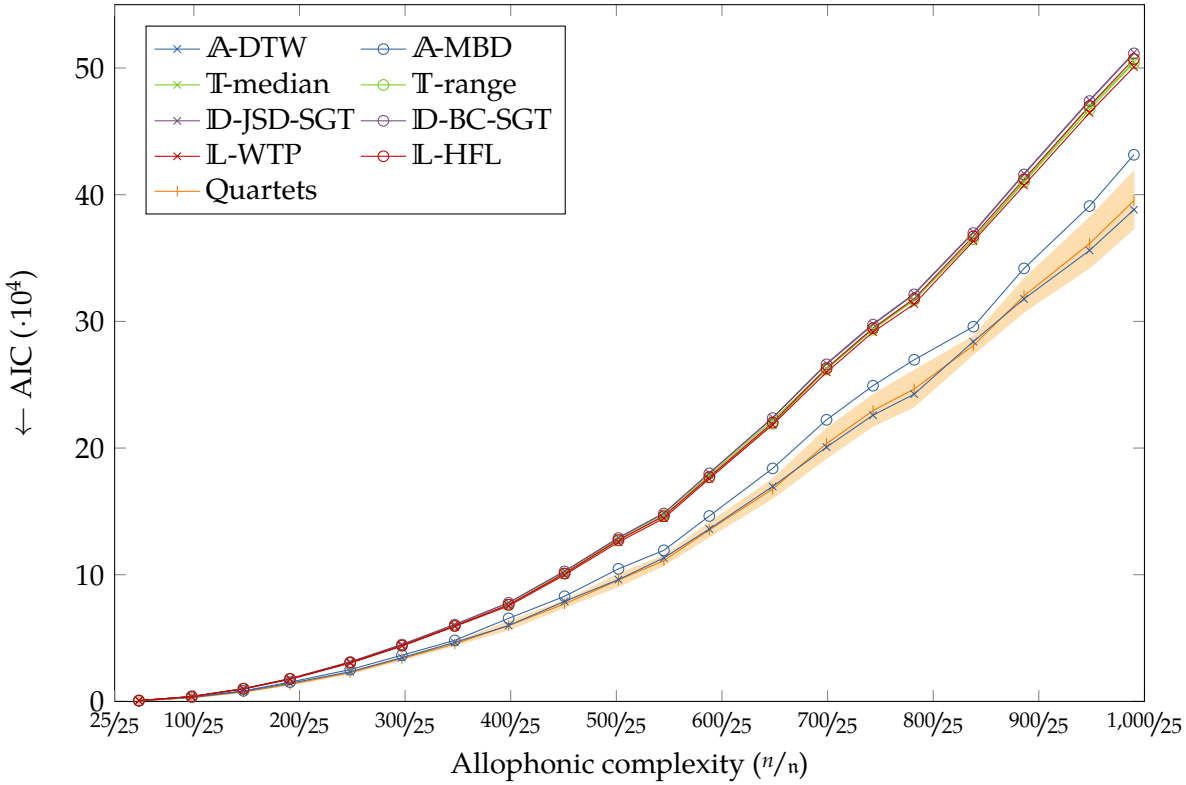


Figure 5.4 — AIC of the flat-response multinomial logistic regression models on the $(\mathfrak{n}+1)$-ary classification task. For quartet models, the solid line marks the average AIC, and the band is bounded by the lowest and the highest values for each allophonic complexity.

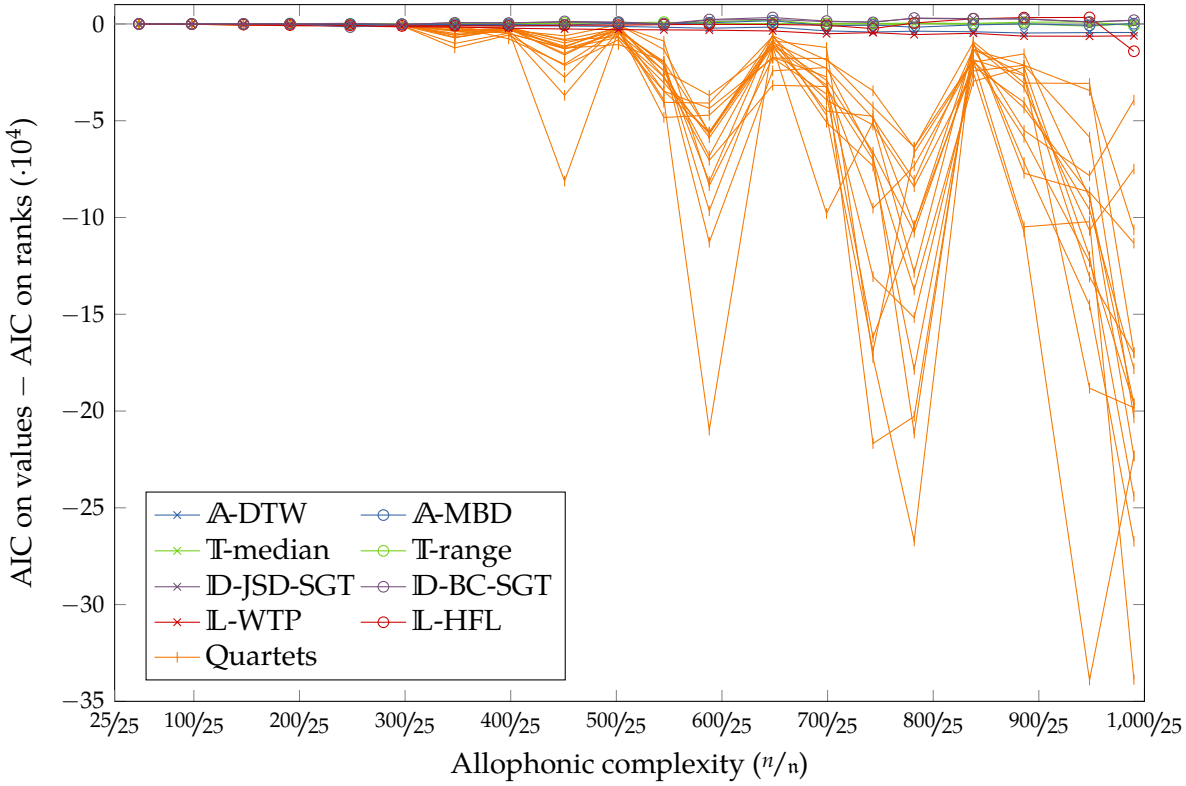


Figure 5.5 — Relative influence of range- and ranks-based standardization on the goodness of fit of the flat-response multinomial logistic regression models of allophony, as a function of allophonic complexity. The gray line denotes the identity function.

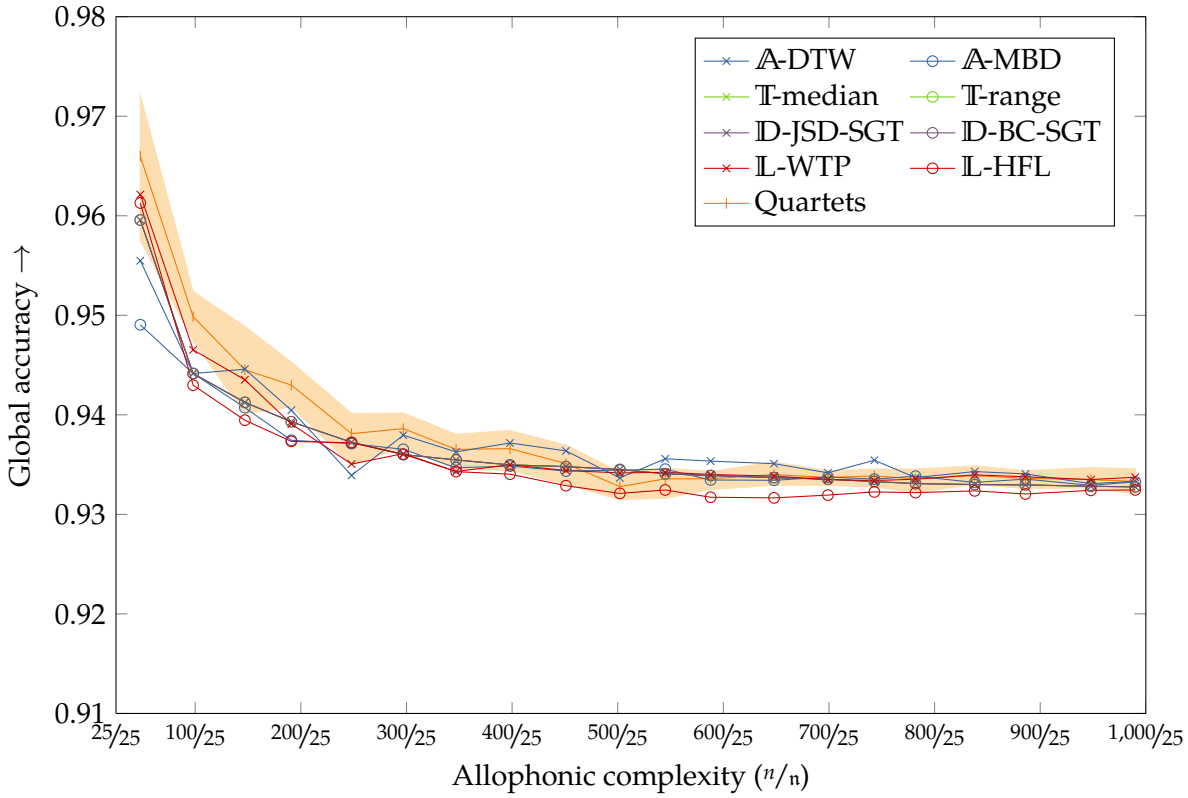


Figure 5.6 — Accuracy of the flat-response multinomial logistic regression models for the $(\mathfrak{n}+1)$-ary classification task. For quartet models, the solid line marks the average AIC, and the band is bounded by the lowest and the highest values for each allophonic complexity.

Table 5.2 — Classwise accuracies of the flat-response combined model on the $(\mathfrak{n}+1)$-ary classification task, as a function of allophonic complexity. The color scale ranges from pure red for accuracies equal to 0 to pure green for accuracies equal to 1. Dots indicate that the corresponding phonemes were not represented at the corresponding allophonic complexities.

n	a	e	i	o	u	aː	eː	iː	oː	uː	w	j	b	d	g	p	t	k	m	n	ɴ	s	z	r	h	⊗
48	.79	.26	.36	0	.	.	.	.	.	.	.	.	.	.	.	.	0	.36	.	0	.	.	.	.	.	1
98	.24	.15	.21	.01	.37	.	0	.	0	.	.	0	.	.	.	.	0	0	0	0	.	.16	.	.	.	1
147	.40	.08	.09	.03	.09	.	0	.	0	0	.	0	.	.	.	.	.02	.05	0	0	0	.12	.	.	.	1
191	.46	.09	.09	.06	.30	.	0	.	0	0	.	0	.	0	0	.	.02	.04	0	0	0	.17	.	.	.	.99
248	.45	.01	.04	.08	.10	0	0	.	0	0	.	0	.	0	0	.	.01	.02	0	0	0	.04	.	0	.	.99
297	.36	.01	.05	.05	.08	0	0	.	0	0	.	0	.	0	0	.	.01	.04	0	0	0	.04	0	0	0	.99
347	.36	0	.05	.05	.11	0	0	.	0	0	.	0	.	0	0	.	.01	0	0	0	0	.07	0	0	0	.99
398	.32	0	.07	.02	.11	0	0	.	0	0	.	0	.	0	0	.	.01	.06	0	0	0	0	0	0	0	1
451	.35	0	.05	.03	.17	0	0	.	0	0	.	0	.	0	0	.	.01	.06	0	0	0	.07	0	0	0	.99
502	.41	0	.04	.03	.11	0	0	.	0	0	.	0	.	0	0	.	.01	.07	0	0	0	.07	0	0	0	.99
545	.38	0	.03	.04	.11	0	0	.	0	0	0	0	0	0	0	.	.01	.08	0	0	0	.08	0	0	0	.99
588	.38	0	.04	.03	.11	0	0	.	0	0	0	0	0	0	0	.	.01	.07	0	0	0	0	0	0	0	.99
648	.40	0	.04	.04	.11	0	0	.	0	0	0	0	0	0	0	.	.01	.07	0	0	0	0	0	0	0	.99
699	.40	0	.05	.02	.11	0	0	.	0	0	0	0	0	0	0	.	.03	.08	0	0	0	0	0	0	0	.99
743	.38	0	.05	.04	.06	0	0	.	0	0	0	0	0	0	0	.	.02	.08	0	0	0	0	0	0	0	.99
782	.37	0	.05	.03	.06	0	0	.	0	0	0	0	0	0	0	.	.03	.08	0	0	0	0	0	0	0	.99
838	.34	0	.04	.04	.06	0	0	.	0	0	0	0	0	0	0	.	.02	.05	0	0	0	.01	0	0	0	.99
886	.33	0	.04	.03	.07	0	0	.	0	0	0	0	0	0	0	.	.02	.04	0	0	0	.01	0	0	0	.99
948	.35	0	.03	.04	.07	0	0	.	0	0	0	0	0	0	0	.	.01	.07	0	0	0	.02	0	0	0	.99
990	.32	0	.03	.04	.11	0	0	.	0	0	0	0	0	0	0	0	.01	.05	0	0	0	.02	0	0	0	.99

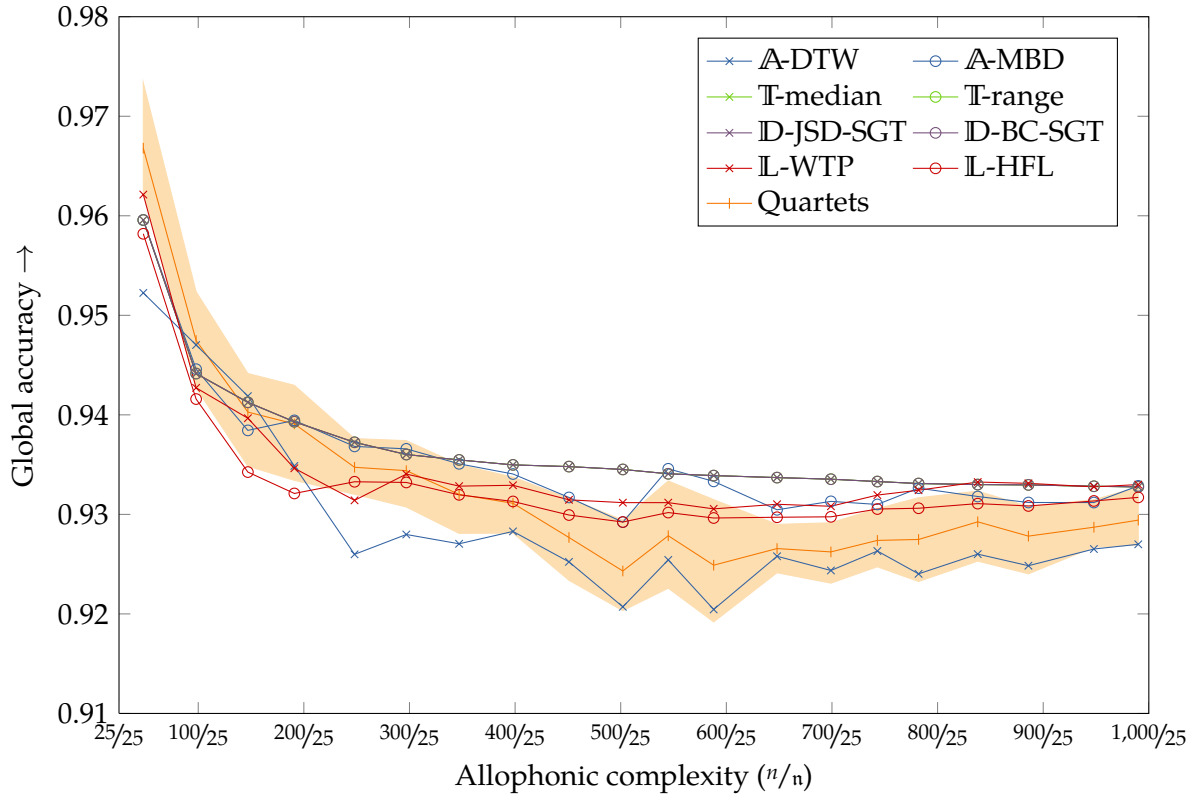


Figure 5.7 — Accuracy of the nested-response multinomial logistic regression models for (n+1)-ary classification task. For quartet models, the solid line marks the average AIC, and the band is bounded by the lowest and the highest values for each allophonic complexity.

Table 5.3 — Classwise accuracies of the nested-response combined model on the (n+1)-ary classification task, as a function of allophonic complexity. The color scale ranges from pure red for accuracies equal to 0 to pure green for accuracies equal to 1. Dots indicate that the corresponding phonemes were not represented at the corresponding allophonic complexities.

n	a	e	i	o	u	aː	eː	iː	oː	uː	w	j	b	d	g	p	t	k	m	n	ɴ	s	z	r	h	⊗
48	.82	0	.36	0	.	.	.	.	.	.	.	.	.	.	.	.	0	.64	.	0	.	.	.	.	.	.99
98	.37	.15	.21	0	.09	.	0	.	0	.	.	0	.	.	.	.	0	0	0	0	.	.13	.	.	.	1
147	.51	.06	.07	.05	.37	.	0	.	0	0	.	0	.	.	.	.	.01	.05	0	0	0	.09	.	.	.	.99
191	.55	.09	.09	.06	.30	.	0	.	0	0	.	0	.	0	0	.	.02	.03	0	0	0	.16	.	.	.	.99
248	.52	.02	.08	.10	.10	0	0	.	0	0	.	0	.	0	0	.	.02	.05	0	0	0	.02	.	0	.	.98
297	.43	.02	.07	.07	.11	0	0	.	0	0	.	0	.	0	0	.	.02	.01	0	0	0	0	0	0	0	.99
347	.45	0	.06	.05	.11	0	0	.	0	0	.	0	.	0	0	.	.01	0	0	0	0	0	0	0	0	.98
398	.44	.01	.07	.02	.11	0	0	.	0	0	.	0	.	0	0	.	.01	.06	0	0	0	0	0	0	0	.98
451	.45	0	.06	.04	.11	0	0	.	0	0	.	0	.	0	0	.	.01	.06	0	0	0	.07	0	0	0	.98
502	.50	0	.06	.04	.14	0	0	.	0	0	.	0	.	0	0	.	.01	.07	0	0	0	.06	0	0	0	.98
545	.45	0	.04	.05	.11	0	0	.	0	0	0	0	0	0	0	.	.01	.06	0	0	0	.06	0	0	0	.98
588	.48	0	.05	.04	.11	0	0	.	0	0	0	0	0	0	0	.	.01	.08	0	0	0	0	0	0	0	.98
648	.46	0	.04	.05	.11	0	0	.	0	0	0	0	0	0	0	.	.01	.08	0	0	0	0	0	0	0	.98
699	.47	0	.06	.02	.11	0	0	.	0	0	0	0	0	0	0	.	.01	.08	0	0	0	0	0	0	0	.98
743	.46	0	.06	.05	.06	0	0	.	0	0	0	0	0	0	0	.	.01	.08	0	0	0	0	0	0	0	.98
782	.44	0	.06	.04	.06	0	0	.	0	0	0	0	0	0	0	.	.01	.07	0	0	0	0	0	0	0	.98
838	.42	0	.05	.05	.07	0	0	.	0	0	0	0	0	0	0	.	.01	.06	0	0	0	.01	0	0	0	.98
886	.42	0	.04	.04	.06	0	0	.	0	0	0	0	0	0	0	.	.01	.05	0	0	0	.01	0	0	0	.98
948	.45	0	.03	.05	.08	0	0	.	0	0	0	0	0	0	0	.	.01	.06	0	0	0	.02	0	0	0	.98
990	.42	0	.04	.04	.12	0	0	.	0	0	0	0	0	0	0	0	.01	.07	0	0	0	.01	0	0	0	.98

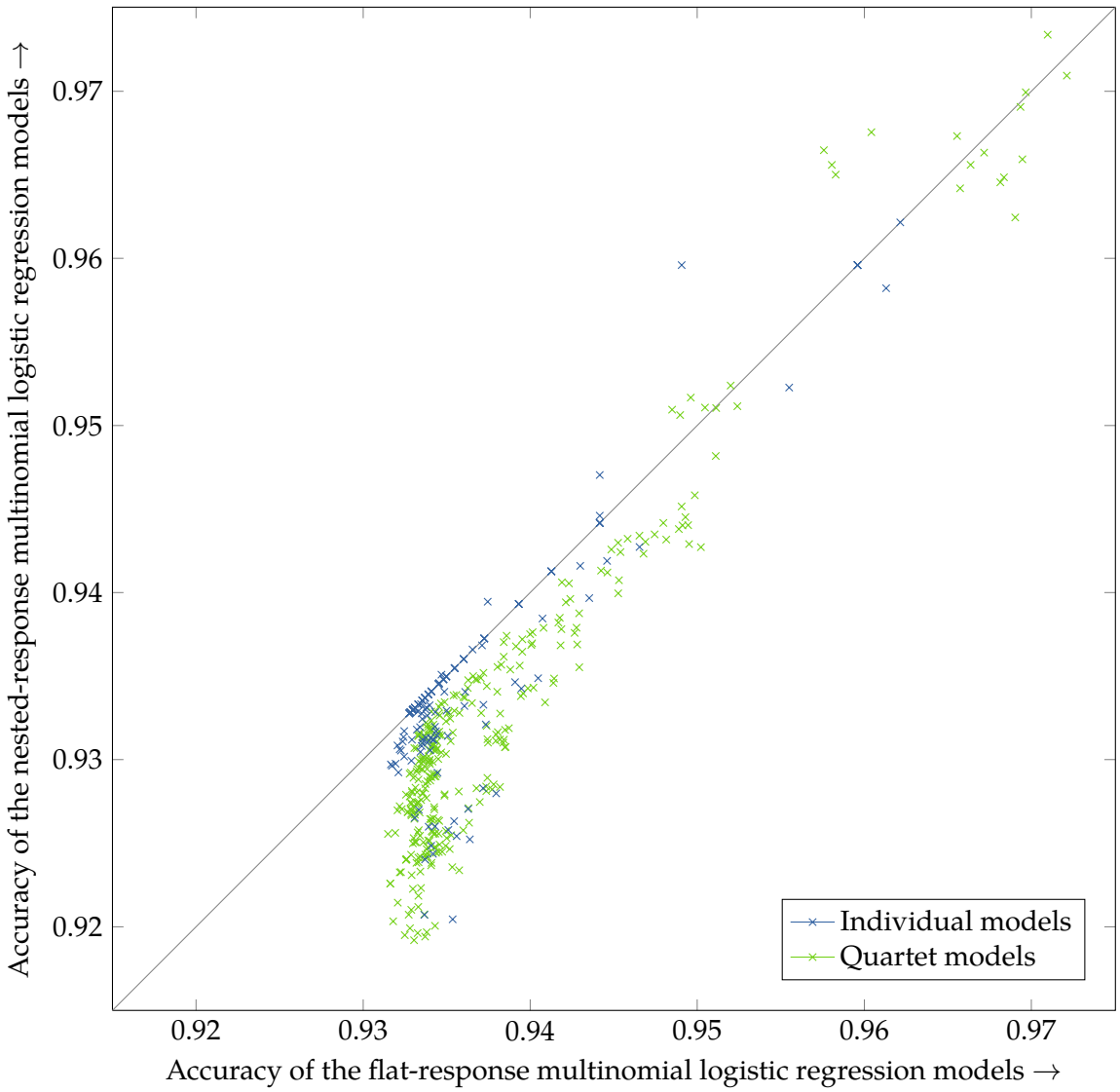


Figure 5.8 — Relative performance of flat- and nested-response multinomial logistic regression models of pairwise phonemehood, across all allophonic complexities. The gray line denotes the identity function.

indicator. In other words, we showed that ranks-based standardization results in a significant loss of information about phonemehood. For this reason, only range-standardized indicators will be considered from this point onwards.

Classification performance For both individual and combined indicators, multinomial logistic regression models appear to be very effective at classifying phone pairs according to their phoneme-like status. Similar to what we observed for binomial models, the global accuracy of multinomial models at any allophonic complexity is strictly greater than 92%, as illustrated in Figures 5.6 and 5.7. Whereas Figure 4.22 suggests that combining indicators yields a significant increase in the accuracy of the predictions, the same cannot be said for multinomial models. For all individual and quartet models, the accuracy of their predictions appears to plateau from $^{500}/_{25}$ allophones per phoneme onwards, meaning that the models do not suffer an increase in allophonic complexity after this point.

As anticipated, nested-response logistic regression multinomial models are globally not as effective as flat-response models. This observation is especially true for quartet models, as illustrated by the wider orange band in Figure 5.7 than in Figure 5.6. Although both approaches

to phonemehood yield highly accurate predictions, flat-response models yield slightly more accuracte predictions than the corresponding nested-response models—regardless of the allophonic complexity of their input, as illustrated in Figure 5.8. Though expected, this result is disappointing, as it means that computational models that fit the true structure of a problem better do not necessarily have a better fit of the data.

The winner takes it all Although informative, contingency tables are space-consuming representations of a given model's predictions. For the sake of brevity, we will thus focus on a single, optimal quartet model in the remainder of the present study. Two quartet models are returned when performing a skyline query on all multinomial logistic regression models based on their respective AIC values across all allophonic complexities—similar to the skyline query described in Section 4.4.2, albeit in reverse order as the AIC is to be minimized:

— $\mathbb{A}$-DTW $\times$ $\mathbb{T}$-range $\times$ $\mathbb{D}$-BC-SGT $\times$ $\mathbb{L}$-WTP;
— $\mathbb{A}$-DTW $\times$ $\mathbb{T}$-median $\times$ $\mathbb{D}$-BC-SGT $\times$ $\mathbb{L}$-WTP.

Interestingly, it is worth noting that these quartet models only differ by their temporal indicator, thus strengthening the optimality of the other three indicators. Among both quartet models, we arbitrarily chose to focus on the latter, which we will refer to as the *combined model* from this point onwards.

Classwise accuracy Classwise accuracies of the flat- and nested-response combined models are reported in Tables 5.2 and 5.3 across all allophonic complexities. Notwithstanding the aforementioned problem of hapax phenomena, the only phoneme-like classes whose classwise accuracy is not always zero are /a/, /e/, /i/, /o/, /u/, /t/, /k/, /s/, and $\otimes$. The figures in these tables confirm that the very high global accuracies observed in Figures 5.6 and 5.7 are mostly due to the almost perfect classification as $\otimes$ of the numerous non-allophonic pairs. Even in this tentative multiclass extension of Peperkamp et al.'s (2006) framework, most—if not all—correct predictions are due to the discrimination of allophonic from non-allophonic pairs. Our results are thus somewhat inconclusive with regards to the question of determining if indicators of allophony are also indicators of phonemehood.

Nonetheless, it is interesting to note that the 8 phonemes for which stricly positive accuracies are observed are not random phonemes, but the 8 most frequent phonemes in the CSJ, as illustrated in Figure 3.2. These globally mediocre results should be put into perspective by the relatively good classwise accuracy for some of these phonemes at the lowest allophonic complexities—and especially in the case of nested-response models. For instance, 79% and 82% of the realizations of /a/ were correctly classified at $^{48}/_{25}$ allophones per phoneme by the flat- and nested-response models, respectively. Not only is /a/ the most frequent phoneme in the corpus—with 212,555 sound tokens—it is also the phoneme with the highest number of allophones at $^{48}/_{25}$ allophones per phoneme—viz. 10 allophones, that is to say, more than twice as many as any other phoneme at this complexity. These additional observations of a frequency effect corroborates our prior hypothesis that a given phoneme's relative frequency and salience may be correlated.

Contingency tables For the sake of completeness, we report the contingency tables obtained for the flat- and nested-response combined model at minimal and maximal allophonic complexities in Tables 5.4, 5.5, 5.6, and 5.7, respectively.

In each cell of a contingency table, the figure indicates the global percentage of (weighted) phone pairs belonging to the column phoneme-like category that were classified as belonging to the row phoneme-like category. Other symbols in the cells read as follows: a dot indicates that the corresponding, grayed-out phoneme does not appear in any phone pair (i.e. it is an hapax phonomenon), a dash marks a cell whose count is equal zero, and an ε marks a cell whose count is stricly positive, yet represents less than a thousandth of the total number of phone pairs.

Table 5.4 — Contingency table for the combined multinomial logistic regression model at $^{48}/_{25}$ allophones per phoneme. Components may not sum to totals because of rounding.

↿	a	e	i	o	u	aː	eː	iː	oː	uː	w	j	b	d	g	p	t	k	m	n	ɴ	s	z	r	h	⊗
a	1.2	0.2	–	0.6	.	.	.	.	.	.	.	.	.	.	.	.	0.2	0.2	.	–	.	.	.	.	.	0.4
e	ɛ	0.1	–	–	.	.	.	.	.	.	.	.	.	.	.	.	–	–	.	–	.	.	.	.	.	–
i	–	–	0.2	–	.	.	.	.	.	.	.	.	.	.	.	.	–	–	.	–	.	.	.	.	.	–
o	–	–	–	0	.	.	.	.	.	.	.	.	.	.	.	.	–	–	.	–	.	.	.	.	.	–
u	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
aː	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
eː	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
iː	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
oː	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
uː	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
w	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
j	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
b	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
d	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
g	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
p	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
t	–	–	–	–	.	.	.	.	.	.	.	.	.	.	.	.	0	–	.	–	.	.	.	.	.	–
k	ɛ	–	–	–	.	.	.	.	.	.	.	.	.	.	.	.	–	0.1	.	0.1	.	.	.	.	.	–
m	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
n	–	–	–	–	.	.	.	.	.	.	.	.	.	.	.	.	–	–	.	0	.	.	.	.	.	–
ɴ	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
s	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
z	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
r	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
h	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
⊗	0.3	ɛ	0.4	0.2	.	.	.	.	.	.	.	.	.	.	.	.	0.2	–	.	–	.	.	.	.	.	95.6

Table 5.5 — Contingency table for the combined nested-response logistic regression model at $^{48}/_{25}$ allophones per phoneme. Components may not sum to totals because of rounding.

↿	a	e	i	o	u	aː	eː	iː	oː	uː	w	j	b	d	g	p	t	k	m	n	ɴ	s	z	r	h	⊗
a	1.3	0.3	–	0.6	.	.	.	.	.	.	.	.	.	.	.	.	0.2	0.1	.	–	.	.	.	.	.	0.5
e	ɛ	0	–	–	.	.	.	.	.	.	.	.	.	.	.	.	–	–	.	–	.	.	.	.	.	–
i	–	–	0.2	–	.	.	.	.	.	.	.	.	.	.	.	.	–	–	.	–	.	.	.	.	.	–
o	–	–	–	0	.	.	.	.	.	.	.	.	.	.	.	.	–	–	.	–	.	.	.	.	.	–
u	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
aː	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
eː	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
iː	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
oː	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
uː	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
w	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
j	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
b	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
d	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
g	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
p	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
t	–	–	–	–	.	.	.	.	.	.	.	.	.	.	.	.	0	–	.	–	.	.	.	.	.	–
k	0.1	–	–	–	.	.	.	.	.	.	.	.	.	.	.	.	–	0.2	.	0.1	.	.	.	.	.	–
m	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
n	–	–	–	–	.	.	.	.	.	.	.	.	.	.	.	.	–	–	.	0	.	.	.	.	.	–
ɴ	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
s	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
z	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
r	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
h	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
⊗	0.2	–	0.4	0.2	.	.	.	.	.	.	.	.	.	.	.	.	0.2	–	.	–	.	.	.	.	.	95.4

Table 5.6 — Contingency table for the combined multinomial logistic regression model at 990/25 allophones per phoneme. Components may not sum to totals because of rounding.

↾	a	e	i	o	u	aː	eː	iː	oː	uː	w	j	b	d	g	p	t	k	m	n	ɴ	s	z	r	h	⊗
a	0.6	0.1	0.1	0.1	ε	ε	ε	.	ε	ε	–	ε	ε	ε	ε	–	0.1	0.1	ε	ε	ε	0.1	ε	0.1	ε	0.6
e	ε	ε	ε	ε	–	–	–	.	–	–	–	–	–	–	–	–	–	–	–	–	–	ε	–	–	–	ε
i	ε	ε	ε	ε	ε	ε	ε	.	ε	ε	ε	ε	–	ε	ε	–	ε	ε	ε	ε	ε	ε	ε	–	ε	ε
o	ε	ε	ε	ε	–	–	ε	.	ε	ε	ε	–	–	ε	–	–	ε	ε	ε	ε	ε	ε	ε	ε	–	ε
u	–	–	–	–	ε	–	–	.	–	–	–	–	–	–	–	–	ε	ε	–	–	–	–	–	–	–	–
aː	–	–	–	–	–	0	–	.	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
eː	–	–	–	–	–	–	0	.	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
iː	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
oː	–	–	–	–	–	–	–	.	0	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
uː	–	–	–	–	–	–	–	.	–	0	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
w	–	–	–	–	–	–	–	.	–	–	0	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
j	–	–	–	–	–	–	–	.	–	–	–	0	–	–	–	–	–	–	–	–	–	–	–	–	–	–
b	–	–	–	–	–	–	–	.	–	–	–	–	0	–	–	–	–	–	–	–	–	–	–	–	–	–
d	–	–	–	–	–	–	–	.	–	–	–	–	–	0	–	–	–	–	–	–	–	–	–	–	–	–
g	–	–	–	–	–	–	–	.	–	–	–	–	–	–	0	–	–	–	–	–	–	–	–	–	–	–
p	–	–	–	–	–	–	–	.	–	–	–	–	–	–	–	0	–	–	–	–	–	–	–	–	–	–
t	–	ε	ε	ε	ε	–	–	.	ε	–	–	–	–	ε	–	–	ε	ε	–	–	ε	–	–	–	–	ε
k	ε	ε	ε	ε	ε	–	–	.	ε	–	–	–	–	–	–	–	ε	ε	ε	ε	ε	ε	ε	–	ε	–
m	–	–	–	–	–	–	–	.	–	–	–	–	–	–	–	–	ε	ε	0	–	–	–	–	–	–	–
n	–	–	–	–	–	–	–	.	–	–	–	–	–	–	–	–	–	ε	–	0	–	–	–	–	–	–
ɴ	–	–	–	–	–	–	–	.	–	–	–	–	–	–	–	–	–	–	–	–	0	–	–	–	–	–
s	ε	ε	ε	ε	–	–	–	.	ε	–	–	ε	–	ε	–	–	ε	ε	ε	ε	–	ε	–	ε	–	–
z	–	–	–	–	–	–	–	.	–	–	–	–	–	–	–	–	–	–	–	–	–	–	0	–	–	–
r	–	–	–	–	–	–	–	.	–	–	–	–	–	–	–	–	–	–	–	–	–	ε	–	ε	–	–
h	–	–	–	–	–	–	–	.	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	0	–
⊗	1.2	0.3	0.6	0.8	0.2	–	ε	.	ε	ε	–	ε	ε	ε	ε	ε	0.5	0.2	0.1	0.2	0.1	0.2	ε	ε	ε	92.6

Table 5.7 — Contingency table for the combined nested-response logistic regression model at 990/25 allophones per phoneme. Components may not sum to totals because of rounding.

↾	a	e	i	o	u	aː	eː	iː	oː	uː	w	j	b	d	g	p	t	k	m	n	ɴ	s	z	r	h	⊗
a	0.7	0.1	0.1	0.2	ε	ε	ε	.	ε	ε	–	ε	ε	ε	ε	–	0.2	0.2	ε	0.1	ε	0.1	ε	0.1	ε	1.3
e	ε	ε	ε	ε	–	–	–	.	–	–	–	–	–	–	–	–	–	–	–	–	–	ε	–	–	–	ε
i	ε	ε	ε	ε	ε	ε	ε	.	ε	ε	ε	ε	–	ε	ε	–	ε	ε	ε	ε	ε	ε	ε	ε	ε	ε
o	ε	ε	ε	ε	–	–	ε	.	ε	ε	ε	–	–	ε	–	–	ε	ε	ε	ε	ε	ε	ε	ε	–	0.1
u	–	–	–	–	ε	–	–	.	–	–	–	–	–	–	–	–	ε	ε	–	–	–	–	–	–	–	–
aː	–	–	–	–	–	0	–	.	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
eː	–	–	–	–	–	–	0	.	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
iː	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
oː	–	–	–	–	–	–	–	.	0	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
uː	–	–	–	–	–	–	–	.	–	0	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
w	–	–	–	–	–	–	–	.	–	–	0	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
j	–	–	–	–	–	–	–	.	–	–	–	0	–	–	–	–	–	–	–	–	–	–	–	–	–	–
b	–	–	–	–	–	–	–	.	–	–	–	–	0	–	–	–	–	–	–	–	–	–	–	–	–	–
d	–	–	–	–	–	–	–	.	–	–	–	–	–	0	–	–	–	–	–	–	–	–	–	–	–	–
g	–	–	–	–	–	–	–	.	–	–	–	–	–	–	0	–	–	–	–	–	–	–	–	–	–	–
p	–	–	–	–	–	–	–	.	–	–	–	–	–	–	–	0	–	–	–	–	–	–	–	–	–	–
t	–	ε	ε	ε	–	–	–	.	ε	–	–	–	–	–	–	–	ε	ε	–	–	ε	–	–	–	–	–
k	ε	ε	ε	ε	ε	–	–	.	ε	–	–	–	–	ε	–	–	ε	ε	ε	ε	ε	ε	ε	–	ε	–
m	–	–	–	–	–	–	–	.	–	–	–	–	–	–	–	–	–	ε	0	–	–	–	–	–	–	–
n	–	–	–	–	–	–	–	.	–	–	–	–	–	–	–	–	–	–	–	0	–	–	–	–	–	–
ɴ	–	–	–	–	–	–	–	.	–	–	–	–	–	–	–	–	–	–	–	–	0	–	–	–	–	–
s	ε	ε	ε	ε	–	–	–	.	ε	–	–	ε	–	ε	–	–	ε	ε	ε	ε	–	ε	–	ε	–	–
z	–	–	–	–	–	–	–	.	–	–	–	–	–	–	–	–	–	–	–	–	–	–	0	–	–	–
r	–	–	–	–	–	–	–	.	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	0	–	–
h	–	–	–	–	–	–	–	.	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	0	–
⊗	1.0	0.3	0.6	0.8	0.2	–	ε	.	ε	ε	–	ε	ε	ε	ε	ε	0.4	0.2	ε	0.2	0.1	0.1	ε	ε	ε	91.9

Finally, strictly positive off-diagonal elements (wrong classifications) are in red, and strictly positive on-diagonal elements (accurate classifications) are in green.

All four tables corroborate our prior observations that multinomial logistic regression models are overwhelmed by non-allophonic pairs, and that the only phonemes whose performance does not appear to be due to chance are the most frequent ones. Indeed, even at maximal allophonic complexity, correctly rejected non-allophonic pairs account for approximately 92% of all predictions. Furthermore, the contingency tables for the nested-response model give an additional example of a frequency effect: the only phonemes as which pairs of phones were wrongfully classified are the aforementioned most frequent phonemes in Japanese, i.e. /a/, /e/, /i/, /o/, /u/, /t/, /k/, and /s/. In order to allow for a better discovery of the phonemic inventory, it thus appears necessary to minimize the impact of non-allophonic pairs.

5.3 Overall assessment

In this chapter, we addressed the limitations of Peperkamp et al.'s (2006) framework for the discovery of allophony and phonemehood. We showed in Section 5.1 that formal and empirical limitations arise from its pairwise formulation. On one hand, Peperkamp et al.'s framework leaves unspecified the question of the transitivity of allophony. On the other hand, it cannot accomodate low-variability phonemes or—to say the least—phonemes that received a single allophone in the quantized representation of the input.

We then carried out various classification experiments whose results unambiguously indicate that the pairwise formulation strongly impedes the discovery of the phonemic inventory of the target language. For instance, we showed in Section 5.2 that, whereas classifying all possible phone pairs in a given allophonic inventory appears to be quite an easy task, the high performance in predictions we observed is mostly due to the correct $\otimes$-rejection of non-allophonic pairs. Moreover, as we have argued since Chapter 4, non-allophonic pairs are simply the spurious byproduct of a combinatioral enumeration as, in the perspective of the discovery of phonemes, only allophonic pairs are informative.

For these reasons, the experiments to be presented in the next chapter consist in a complete reformulation of Peperkamp et al.'s (2006) framework so that individual phones—rather than pairs of phones—become the elementary objects to be manipulated. In order to do so, we will create a new representation for the input data, and recast the task of acquiring phonemes as a standard partitioning-clustering problem.

CHAPTER 6
PHONEMEHOOD REDUX

In this chapter, we propose a complete reformulation of Peperkamp et al.'s (2006) framework so that individual phones—rather than pairs of phones—become the elementary objects to be manipulated. We indeed showed in the preceding chapters of this study that Peperkamp et al.'s framework suffers from the considerable number and smeared distribution of non-allophonic pairs, for all (combinations of) indicators of allophony and at any allophonic complexity. Moreover, we have argued that non-allophonic pairs are ultimately irrelevant for the purpose of discovering the phonemic inventory of the target language as they are simply the spurious byproduct of a combinatorial enumeration. For these reasons, we shift the primitive data structure from a pair of phones to a single phone. To this aim, we use the phone-to-phone dissimilarity ratings provided by (a combination of) indicators of allophony to define points in a Euclidean space so that each phone in the allophonic inventory is represented by a point in the Euclidean space, and that the point-to-point distances correspond to the phone-to-phone dissimilarities. Under such a representation, the phonemic inventory of the target language can eventually be learned directly from a partition of a given allophonic inventory whereby each subset in the partition is to be interpreted as a putative phoneme.

This chapter is divided into six main sections. In Section 6.1, we present the fundamental principles of three-way multidimensional scaling, the statistical technique we used to compute a novel, pair-free representation for each indicator of allophony (or combination thereof). In Section 6.2, we present a first assessment of indicators' phoneme-wise informativeness in terms of prognoses of phonemehood. Various classification and clustering experiments are then reported in Sections 6.3, 6.4, and 6.5. Finally, Section 6.6 contains a general assessment of the work presented in this chapter.

6.1 Shifting the primitive data structure

As we argued in Section 5.1.1, phonemehood is a polyadic relation that exhaustively describes the realizations of each and every phoneme in the target language, whereas allophony is a binary relation that merely indicates whether two phones are realizations of the same phoneme but, in that event, contains no information regarding the identity of the phoneme of which they both are realizations. Furthermore, we showed that theoretical and empirical limitations arise (transitivity and hapax phonomena, respectively) when the first step in the discovery of the phonemic inventory of a given language consists in learning the allophony relation between all possible pairs of phones. These limitations are due to the pairwise nature of Peperkamp et al.'s (2006) framework. For these reasons, our goal in this section is to define a new representation of the input data whereby the primitive data structure to be considered by the learner is not a pair of phones, but a single phone.

Unpairing phone pairs In Chapter 4, we showed that the phone-to-phone dissimilarity ratings we refer to as indicators of allophony are indeed correlated with allophony: non-allophonic pairs tend to be more dissimilar than allophonic pairs depending, to a certain extent, on the allophonic complexity of the input as well as the linguistic cue each indicator focuses on. As a consequence, we can reasonably consider that indicators of allophony confirm the aforementioned compactness hypothesis (Duin, 1999; Pękalska et al., 2003) whereby objects that are underlyingly similar (a given phoneme's allophones) are also close in their representation (a given indicator's phone-to-phone dissimilarities). Whereas indicators of allophony have proven to be accurate cues for the discovery of phonemes, they however offer no workable representation for the phones they rely on. To draw a non-technical analogy, examining an indicator of allophony is akin to observing a matrix of city-to-city distances as the crow flies, but not having access to a map. For example, knowing that Paris is closer to Ruffec than to Los Angeles gives—at a first glance—no information on the location of each city. However, the complete matrix of all pairwise distances (e.g. Paris–Ruffec, Paris–Los Angeles, and Ruffec–Los Angeles) can be interpreted as a set of relative constraints on the estimation of each city's location: assuming that we work in a Euclidean 2-dimensional projection of the surface of the Earth (i.e. a common map), those three cities define the vertices of a triangle whose edges' length are given by the pairwise distances between the corresponding cities. In this section, we describe the positioning technique we used to infer such Euclidean embeddings for our allophonic inventories, based on indicators of allophony's dissimilarity ratings.

6.1.1 Multidimensional scaling

The classic Euclidean embedding technique we used is known as *multidimensional scaling* (henceforth MDS; Torgerson, 1952; Groenen & van de Velden, 2004; Borg & Groenen, 2005). Generally speaking, the input data used for MDS consist in dissimilarities between pairs of objects, and the objective of MDS is to represent these dissimilarities as distances between points in a multidimensional Euclidean space such that the distance correspond as closely as possible to the dissimilarities. Following Borg & Groenen (2005), we use MDS as a psychological model that

> ❝[explains] perceived dissimilarity as the result of a mental arithmetic that mimic the distance formula. According to this model, the mind generates an impression of dissimilarity by adding up the perceived differences of the two objects over their properties.❞

In the case at hand, the objective of MDS is to locate n points—i.e. one for each phone—in a low-dimensional Euclidean space in such a way that the distances between the points approximate the $n \times n$ input dissimilarities $\mathbf{\Delta}$ of a given indicator of allophony. It is worth noting that previously defined phone-to-phone dissimilarity matrices satisfy all the requirements of MDS, as they are by definition square, symmetric, non-negative, and hollow (cf. Section 4.1.2).

More precisely, given an a priori dimensionality $q \in \mathbb{N}$ such that $1 \leq q < n - 1$, we want to find an $n \times q$ *metric configuration* matrix $\mathbf{M} \equiv [m_{il}]$ whereby each phone $p_i \in P$ in the allophonic inventory is represented by a point $\mathbf{m}_i \equiv (m_{i1}, \ldots, m_{iq})$ in a q-dimensional Euclidean space (if $q \geq n - 1$, the solution is trivial; Borg & Groenen, 2005). Furthermore, we want the input phone-to-phone dissimilarities to constrain the output point-to-point distances. Concretely, let $d_{ij}(\mathbf{M})$ denote the Euclidean distance between the points $\mathbf{m}_i$ and $\mathbf{m}_j$, given by

$$d_{ij}(\mathbf{M}) \equiv \sqrt{\sum_l (m_{il} - m_{jl})^2} \quad \text{where} \quad l \in (1, 2, \ldots, q), \tag{6.1}$$

we aim at finding the optimal Euclidean embedding of $\mathbf{\Delta}$, i.e. the embedding where the output distances $d_{ij}(\mathbf{M})$ correspond as closely as possible to the input dissimilarities δ_{ij}.

Stress and SMACOF Following de Leeuw & Mair (2009), we make the optimization problem more precise by defining the objective *stress* function $\sigma(\mathbf{M})$. The stress of a given metric configuration $\mathbf{M}$ with regards to the input pairwise dissimilarities $\mathbf{\Delta}$ and the pairwise weights $\mathbf{W}$ is given by the weighted sum of the squared differences between the corresponding input dissimilarities

and output distances (a.k.a. the standard residual sum of squares; Pękalska & Duin, 2005; p. 136). Formally, we have

$$\sigma(\mathbf{M};\boldsymbol{\Delta},\mathbf{W}) \equiv \sum_{i<j} w_{ij}\big(\delta_{ij} - d_{ij}(\mathbf{M})\big)^2. \tag{6.2}$$

It is worth mentioning that, in the case at hand, the previously defined $n \times n$ weight matrices $\mathbf{W}$ also satisfy the assumptions of symmetry, non-negativity, and hollowness (cf. Section 4.3.3).

Because stress values are given by the added differences between corresponding input dissimilarities and output distances, this objective function is to be minimized. The iterative strategy we used to minimize the stress function is referred to in the MDS literature as SMACOF (short for *scaling by majorizing a complicated function*; de Leeuw & Mair, 2009). Although describing the practical details behind this optimization strategy is beyond the scope of this research, it is worth highlighting the fact that SMACOF has the desirable property to guarantee a series of non-increasing stress values with a linear convergence rate (de Leeuw, 1988). Put another way, performing an additional iteration of the SMACOF algorithm can not decrease the quality of the output configuration. However, depending on the number n of objects in the dataset and the composition of the dissimilarity matrix, a large number of iterations may be required if a high accuracy is needed (Groenen & Heiser, 2000; Bronstein et al., 2005; Rosman et al., 2008).

It is worth noting that the optimality of the output metric configuration $\mathbf{M}$ only depends on the input dissimilarities $\boldsymbol{\Delta}$ and the pairwise weights $\mathbf{W}$—no reference data structure such as $\mathfrak{P}$ or $\mathbf{A}^*$ is involved in the computation of the stress function. Though it is not a machine learning technique stricto sensu, MDS is thus to be considered unsupervised or—to say the least—unaware of the true structure of the data it manipulates.

Strike a pose, there's nothing to it! Because Euclidean distances are preserved under rotation, translation, and reflection, these operations may be applied to a given metric configuration without affecting its stress. Extending our prior analogy with geographic distances, the distance between Paris and Los Angeles is not affected by turning the map upside down or, simply, the rotation of the Earth. Most MDS programs rely on this insensitivity to distance-preserving transformations (i.e. isometries) and, for the sake of efficiency, compute metric configurations that are only determined up to location (Venables & Ripley, 2002; R Development Core Team, 2010; cf. `stats::cmdscale` and `mass::isoMDS`). In other words, distinct metric configurations may have equal (and optimal) stress values with regards to a given input dissimilarity matrix. Consequently, it is worth noting that although MDS outputs workable location estimates, individual coordinates and dimensions are not meaningful.

Non-metric SMACOF In Chapter 4, we only constrained indicators of allophony to be symmetric, non-negative dissimilarity measures. Therefore, we did not constrain indicators to be true metrics as neither the identity of indiscernibles (i.e. $\forall\{p_i, p_j\} \subseteq P,\ \delta_{ij} = 0 \Rightarrow p_i = p_j$) nor the triangle inequality (i.e. $\forall\{p_i, p_{i'}, p_j\} \subseteq P,\ \delta_{ij} \leq \delta_{ii'} + \delta_{i'j}$) were enforced. In the perspective of a Euclidean—hence metric—embedding, such dissimilarities are considered to convey qualitative instead of quantitative information (Pękalska & Duin, 2005): whereas dissimilarities convey more information than a mere pairwise ranking (cf. Section 5.2), they can not be used as they are to infer sound location estimates. We can consequently think of transformations of the dissimilarities that—while preserving the pairwise ranking, to the bare minimum—allow for a better fit of the configuration $\mathbf{M}$ in the Euclidean space. If such a transformation f obeys the monotonicity constraint that $\delta_{ij} < \delta_{i'j'} \Rightarrow f(\delta_{ij}) < f(\delta_{i'j'})$, the MDS procedure is referred to as *non-metric* (Shepard, 1962a,b; Borg & Groenen, 2005; de Leeuw & Mair, 2009) and the corresponding non-metric stress function becomes

$$\sigma(\mathbf{M}, f;\boldsymbol{\Delta},\mathbf{W}) \equiv \sum_{i<j} w_{ij}\big(f(\delta_{ij}) - d_{ij}(\mathbf{M})\big)^2. \tag{6.3}$$

In non-metric MDS, the monotonic transformation f is determined and optimized by the program together with the metric configuration $\mathbf{M}$. In our experiments, the optimal monotonic

transformation f was recomputed after each SMACOF iteration (setting the parameter `modulus` to 1 in `smacof`; de Leeuw & Mair, 2009).

Transforming the observed dissimilarity matrix $\mathbf{\Delta}$ calls for a proper treatment of tied values, i.e. when $\delta_{ij} = \delta_{i'j'}$ for two phone pairs $\{p_i, p_j\}$ and $\{p_{i'}, p_{j'}\}$. In that case, two different approaches are commonly suggested in the MDS literature: the unrestrictive (a.k.a. primary) approach that does not necessarily requires that transforming ties yields ties, i.e. $\delta_{ij} = \delta_{i'j'} \not\Rightarrow f(\delta_{ij}) = f(\delta_{i'j'})$, and the restrictive (a.k.a. secondary) approach that does (de Leeuw & Mair, 2009). In the absence of a priori linguistic knowledge that would justify the use of the restrictive approach—we can not interpret ties in a given indicator's values, except as numerical accidents—we chose to consider only the unrestrictive approach in our experiments.

Three-way SMACOF In its classic definition, MDS can only accomodate one input dissimilarity matrix $\mathbf{\Delta}$, and this approach to Euclidean embedding is known as two-way (i.e. phone to phone) MDS. Various extensions of the original MDS procedure have been proposed to infer a single metric configuration from various dissimilarity matrices. In the case at hand, such extensions allow us to combine κ different indicators of allophony $\mathbf{\Delta}_1, \ldots, \mathbf{\Delta}_\kappa$ (typically an acoustic, a temporal, a distributional, and a lexical indicator) into a single metric configuration.

Combining κ separate $n \times n$ dissimilarity matrices into a single $n \times q$ group configuration $\mathbf{Z}$, to be defined hereafter, is known as *MDS for individual differences* and *three-way MDS* (i.e. phone to phone to indicator; Horan, 1969; Borg & Groenen, 2005; de Leeuw & Mair, 2009). The typical setting for three-way MDS arises when $n \times n$ dissimilarity ratings are available from κ different—yet comparable—data sources, e.g. subjects, replications, or indicators of allophony. The objective of three-way MDS is best summarized in the title of Horan's (1969) seminal study:

❝Combining observations when individuals have different perceptual structures.❞

Indeed, the major assumption behind three-way MDS is that individual spaces are systematical distortions of a shared space (MDS(X) User Manual, 1981). Thus, even though the optimization involves the computation of a configuration matrix $\mathbf{M}_k$ for each individual, the standard strategy consists in introducing restrictions on the configurations $\mathbf{M}_k$ in terms of a linear decomposition

$$\mathbf{M}_k \equiv \mathbf{Z}\mathbf{R}_k, \tag{6.4}$$

where $\mathbf{Z} \equiv [z_{il}]$ is the $n \times q$ shared *group configuration* and each $\mathbf{R}_k \equiv [r_{ll'k}]$ is an individual $q \times q$ regression weight matrix to be estimated simultaneously (de Leeuw & Mair, 2009). As emphasized in the MDS(X) User Manual (1981), the group configuration $\mathbf{Z}$ is a compromise between all individuals' configurations, and it may conceivably describe the configuration of no single individual.

In the case of three-way MDS, the stress of the group configuration is given by the sum of the stress values of the individual configurations. Formally, we have

$$\sigma(\mathbf{Z}, \mathbf{R}_1, f_1, \ldots, \mathbf{R}_\kappa, f_\kappa; \mathbf{\Delta}_1, \ldots, \mathbf{\Delta}_\kappa, \mathbf{W}) \equiv \sum_k \sigma(\mathbf{Z}\mathbf{R}_k, f_k; \mathbf{\Delta}_k, \mathbf{W}) \tag{6.5}$$

$$= \sum_{i<j} w_{ij} \left(\sum_k \big(f_k(\delta_{ijk}) - d_{ij}(\mathbf{Z}\mathbf{R}_k)\big)^2 \right) \tag{6.6}$$

It is worth noting that this definition is somewhat simpler than the one given by de Leeuw & Mair (2009; eq. 21, p. 7). This simplification arises from the fact that, in the case at hand, weights are not conditional on the individuals (i.e. all indicators share the same weighting scheme) and, hence, we rely on a single weight matrix $\mathbf{W}$ rather than κ different weight matrices.

All in all, given weighted three-way dissimilarities data, the objective of a three-way MDS procedure is to solve simultaneously for the group configuration $\mathbf{Z}$, the individual regression matrices $\mathbf{R}_1, \ldots, \mathbf{R}_\kappa$, and the monotonic transformations $f_1, \ldots, f_\kappa$, so as to optimize the fit of the group configuration to the (monotonically transformed) dissimilarities data.

Individual differences In order to control to which extent each individual space is a distortion of the underlying group space, we introduce further restrictions in the definition of three-way

MDS. To do so, we constrain the form of the individual regression matrices $\mathbf{R}_1, \ldots, \mathbf{R}_\kappa$ (Horan, 1969; Carroll & Chang, 1970; MDS(X) User Manual, 1981; de Leeuw & Mair, 2009). We consider three standard variations of the abovementioned model, ranging from fully unconstrained to fully constrained.

If no constraint is imposed, i.e. every regression matrix $\mathbf{R}_k$ is a general matrix, the model is known as the *generalized Euclidean model* (a.k.a. IDIOSCAL, short for *individual differences in orientation scaling*) as it relies on the generalized Euclidean distance

$$d_{ij}(\mathbf{M}_k) \equiv d_{ij}(\mathbf{Z}\mathbf{R}_k) \equiv \sqrt{\sum_l \sum_{l'} u_{ll'k}(z_{il} - z_{jl})(z_{il'} - z_{jl'})}, \tag{6.7}$$

where $\mathbf{U}_k \equiv \mathbf{R}_k \mathbf{R}_k^{\mathrm{T}}$. Under this unconstrained model, every individual systematically transforms the group space first by a rotation or a reflection, and then by scaling each dimension.

If every regression matrix $\mathbf{R}_k$ is constrained to be a diagonal matrix, then $\mathbf{U}_k \equiv \mathbf{R}_k \mathbf{R}_k^{\mathrm{T}} = \mathbf{R}_k^2$ is also diagonal. The corresponding MDS model is known as the *weighted Euclidean model* (a.k.a. INDSCAL, short for *individual differences scaling*), and the underlying weighted Euclidean distance can straightforwardly be defined as

$$d_{ij}(\mathbf{M}_k) \equiv d_{ij}(\mathbf{Z}\mathbf{R}_k) \equiv \sqrt{\sum_l r_{llk}^2 (z_{il} - z_{jl})^2}. \tag{6.8}$$

Under this model, scaling becomes the only possible individual transformation. In other words, this model explains differences between individuals by a differential weighting of each dimension by each individual (MDS(X) User Manual, 1981), i.e. all indicators rely on the same underlying group space, but different dimensions of that space may be more or less salient to each indicator.

Finally, if every regression matrix $\mathbf{R}_k$ is further constrained to be an identity matrix, we have $\mathbf{M}_k \equiv \mathbf{Z}\mathbf{R}_k = \mathbf{Z}$. This model, the simplest of all three-way MDS models we consider in this study, is known as the *unweighted Euclidean model* (a.k.a. the *identity model*), and the definition of the distance function can be simplified to the usual Euclidean distance, i.e.

$$d_{ij}(\mathbf{M}_k) \equiv d_{ij}(\mathbf{Z}) \equiv \sqrt{\sum_l (z_{il} - z_{jl})^2} \tag{6.9}$$

Under this very constrained model, no individual transformation is allowed and, hence, individual spaces do not depart from the group space.

In Chapter 4, we computed various (acoustic, temporal, distributional, and lexical) indicators to assess the allophonic dissimilarity of all pairs of phones in a given allophonic inventory. We are however unable not assert that indicators of different classes use the response scale and the underlying (phonological) group space in identical ways. In fact, in light of the recurrent discrepancies in performance we observed in Chapters 4 and 5 between all four indicator classes, we are able to assert that they do not use the group space identically. Our dissimilarity measurements are thus conditional on each indicator: all observations within a dissimilarity matrix are comparable, but not corresponding observations between matrices (a property commonly referred to as *matrix-conditionality* or *split-by-matrices* in the three-way MDS literature; Borg & Groenen, 2005). Whereas, for instance, it would be meaningful to say that a given phone pair is acoustically twice as dissimilar as another phone pair, it would make no sense to say that a given phone pair is acoustically twice as dissimilar as it is distributionally.

For this reason, the unweighted Euclidean model—whereby no individual transformation is possible—appears to be too restrictive for the case at hand, especially compared to the generalized Euclidean (IDIOSCAL) and the weighted Euclidean (INDSCAL) models. The difference between these two models arises from the fact that scaling, rotations, and reflections are permissible in the former, while only scaling is permissible in the latter. As in the case of tied values, we favored an unrestricted approach in our experiments: in the absence of a priori linguistic knowledge that would oppose the use of rotations and reflections in the computation of the optimal Euclidean embedding of an allophonic inventory from three-way dissimilarity data, all MDS configurations to be further examined were computed under the generalized Euclidean model.

Convergence criteria The convergence of SMACOF can be controlled by setting the values of the following parameters: the maximal number of iterations above which optimization is stopped, and the minimal gain in stress between two consecutive iterations below which optimization is stopped (`maxit` and `eps`, respectively, in `smacof`; de Leeuw & Mair, 2009). Contrary to other MDS optimization procedures, the formulation of stress used in SMACOF is dependent on the number of objects to embed, as well as on the scale of the dissimilarities (Groenen & van de Velden, 2004). Therefore, we can not reasonably set a value for the minimal gain in stress between two consecutive iterations that would be appropriate for all indicators and allophonic complexities and, hence, we solely relied on the maximal number of iterations to control the convergence of SMACOF in our experiments.

It is worth emphasizing that computing three-way MDS configurations is computationally prohibitive: in the perspective of embedding n phones in a q-dimensional Euclidean space based on κ indicators, the time and space complexity of a single SMACOF iteration is necessarily in $\mathcal{O}(n^2q\kappa)$. Because we observed that the greatest gains in stress occur during the ten or so first iterations, we chose to control the convergence of SMACOF by limiting—somewhat arbitrarily—the optimization procedure to 100 iterations. Increasing the number of iterations would certainly yield metric configurations with a better fit to the input dissimilarities; however, considering the computation time and power required for a single iteration, we were unable to perform more than 100 iterations for any given complexity and indicator. Whereas, for example, fitting a multinomial logistic regression model is a matter of minutes, fitting the corresponding metric configuration is, on the same computer, a matter of weeks. For this reason, we were unfortunately unable to compute three-way MDS configurations for allophonic complexities above $^{588}/_{25}$ allophones per phoneme.

The curse of dimensionality Before turning to the actual MDS configurations, we need to address the issue of dimensionality. The number of dimensions q of the Euclidean space whererin the MDS procedure embeds the objects is, indeed, to be specified a priori. Whereas guidelines and techniques have been proposed to find the optimal dimensionality in which to embed a given dataset (Venables & Ripley, 2002; Borg & Groenen, 2005), they require to test the goodness of fit of all optimal configurations across a considerable range of dimensions and, hence, are computationally intractable in the case at hand. Moreover, local minima—the pitfall for virtually all optimization problems, whereby convergence is reached while the solution is not globally optimal—can occur in low-dimensional spaces. Groenen & Heiser (1996) indeed showed that, in the words of Borg & Groenen (2005):

> ❝local minima are more likely to occur in low-dimensional solutions (especially unidimensional scaling). For high-dimensional solutions local minima are rather unlikely whereas fulldimensional scaling has no local minimum problem.❞

All things considered, we are in a situation where we have to mitigate the trade-off between the tractability (low dimensionality) and the correctness (high dimensionality) of our experiments, and where the only available methodology is trial and error. To our knowledge, the experiments reported in this chapter are the first attempt at using three-way MDS on language-related data. In the absence of any rule of thumb or comparable scholarship, we thus arbitrarily chose to test the following dimensionalities: $q \in (2, 10, 20, 30)$.

Rejected alternatives It is worth noting that other techniques may have been used to provide location estimates for all phones in a given allophonic inventory, most notably:

— using *alternating least squares scaling* (a.k.a. ALSCAL; Takane et al., 1977): this embedding technique natively accomodates three-way dissimilarities but its implementation is, to our knowledge, now only available as weakly documented legacy FORTRAN code;

— averaging the dissimilarities over the indicators, and scaling the aggregated matrix (Borg & Groenen, 2005): this approach is precluded by the matrix-conditionality of the data at hand, as averaging the dissimilarities would not be meaningful;

— scaling each indicator separately, and combining the different configurations (Borg & Groenen, 2005): this approach leaves unspecified the issue of the combination itself;
— scaling each indicator separately, and learning phonemes in *parallel universes* (Patterson & Berthold, 2001; Wiswedel et al., 2010): although this approach circumvents the issue of the combination by using as many spaces as indicators, few classification or clustering algorithms have been extended to accomodate this framework;
— using a dissimilarity space (Duin & Pękalska, 2012) wherein each phone $p_i \in P$ is represented by the vector $(\delta_{i11}, \ldots, \delta_{in\kappa})$ of the dissimilarities between p_i and all phones in P across all κ indicators: in this approach, the number of dimensions $q = n\kappa$ of the dissimilarity space increases with the allophonic complexity of the input, thus hindering the comparability across complexities.

6.1.2 Visualizing phone configurations

In this section, we present a first examination of the metric configurations we computed for the individual indicators of allophony 𝔸-DTW, 𝕋-median, 𝔻-BC-SGT, 𝕃-WTP, as well as for the group model, i.e. a three-way combination of these indicators.

Stress and dimensionality The stress of the optimal metric configurations obtained for all five models are presented in Figures 6.1, 6.2, 6.3, 6.4, and 6.5 as a function of allophonic complexity and dimensionality. As expected, the more dimensions, the better the Euclidean embedding: the two-dimensional solution is always—and by far—the least optimal embedding of the underlying allophonic inventory. The stress of the optimal configuration, however, is not linearly related to the dimensionality of the embedding space: for all five models, the gain in stress from 20 to 30 dimensions is always lesser than the gain from 10 to 20 and, especially, 2 to 10. Be that as it may, embedding allophonic inventories in 30-dimensional Euclidean space appears to be the optimal solution among the ones we tested. For this reason, all mentions of metric configurations will refer to 30-dimensional configurations from this point onwards.

The stress values of the metric configurations obtained for all five models in 30 dimensions are presented in Figure 6.6. Although the three-way stress values—given here for the sake of completeness—are not comparable to the two-way stress values, it is worth noting that dissimilarities collected in the acoustic and the distributional indicators are more easily embeddable than the home-brewed dissimilarities collected in the temporal and the lexical indicators. Although non-metric MDS is specifically designed to handle arbitrary dissimilarities, these results suggest that a possible perspective for further research would consist in defining indicators of allophony that either satisfy all metric properties or are easily embeddable in a metric space, so that the point-to-point distances really reflect the phone-to-phone dissimilarities.

Two-dimensional configurations For the sake of illustration, the optimal 2-dimensional metric configurations obtained at $^{48}/_{25}$ allophones per phoneme are presented in Figure 6.7. Although we have shown that using 2-dimensional Euclidean spaces yields the metric configurations with the worst fit to the input dissimilarities, 2 is the maximal dimensionality at which visualizations can be conveniently reported in the present (2-dimensional) dissertation. The metric configurations in Figure 6.7 were purposedly plotted against unannotated axes to emphasize the fact that, in such embeddings, only relative distances are meaningful—and not individual coordinates or dimensions. It is worth noting, however, that each scatter plot's aspect ratio is 1:1, so that the plotted distances truly reflect the computed distances.

Unfortunately, no linguistically relevant pattern emerge in these plots: allophones are scattered, and no natural classes appear.

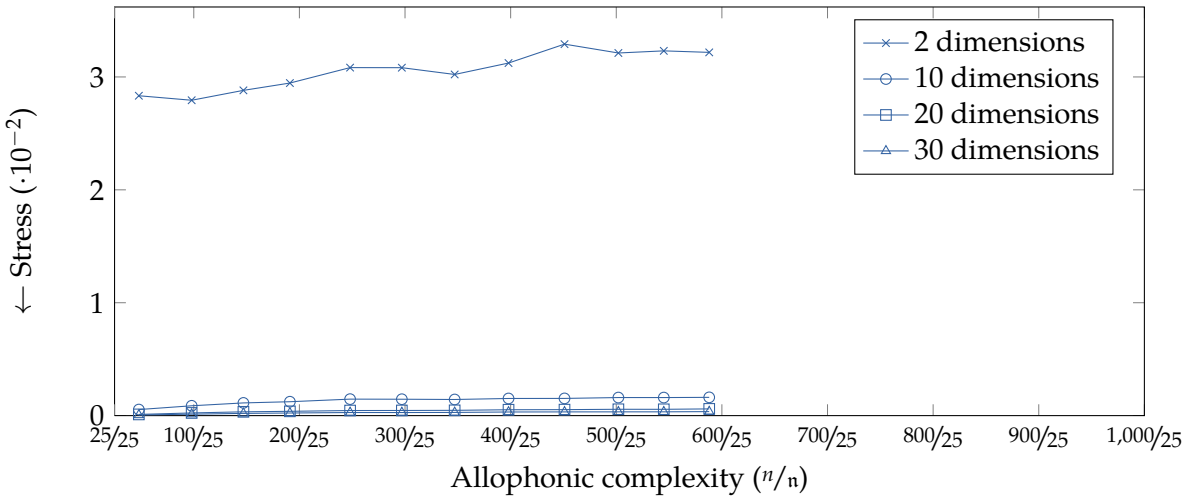


Figure 6.1 — Stress of the optimal metric configurations for 𝔸-DTW, as a function of allophonic complexity and dimensionality.

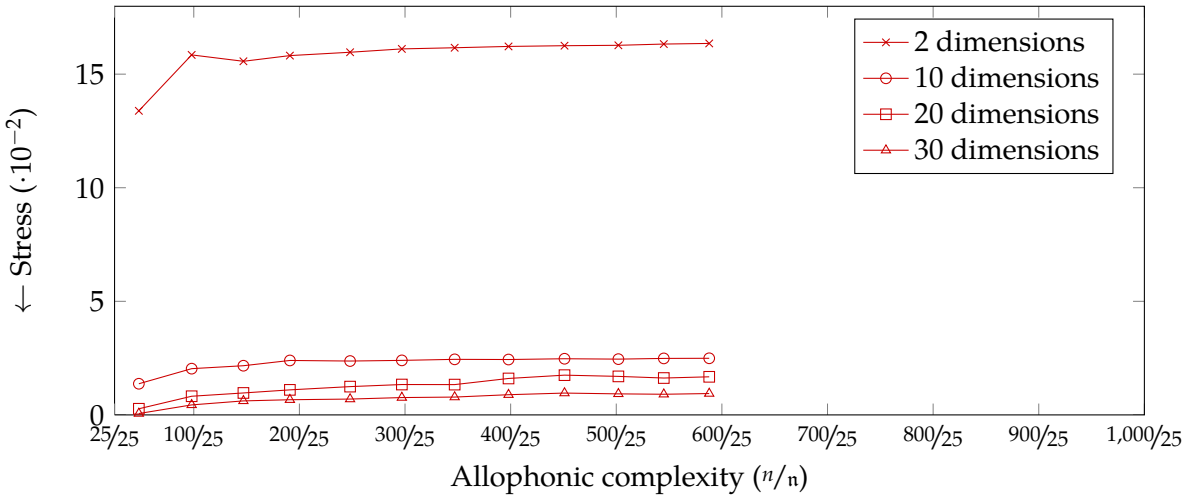


Figure 6.2 — Stress of the optimal metric configurations for 𝕋-median, as a function of allophonic complexity and dimensionality.

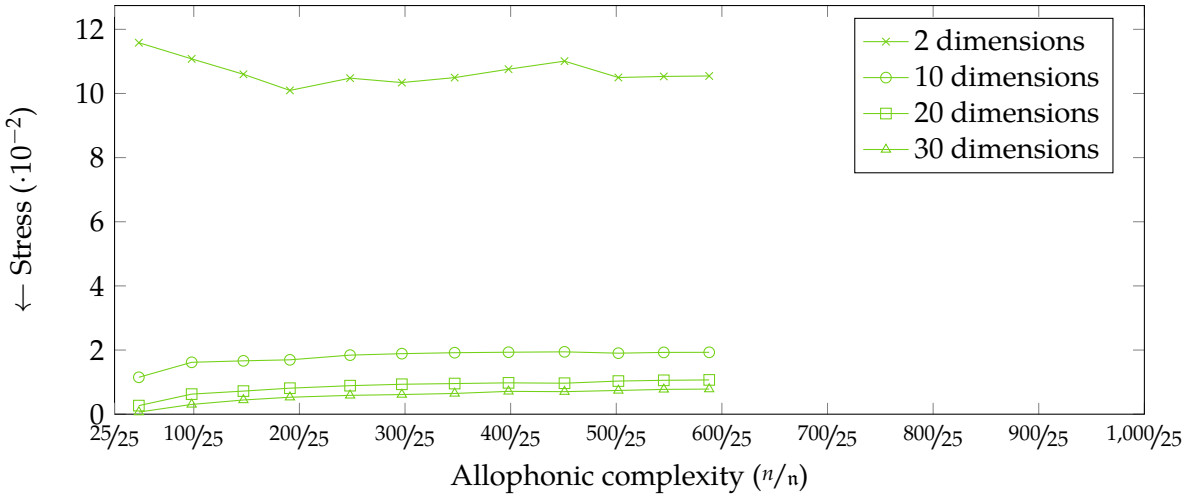


Figure 6.3 — Stress of the optimal metric configurations for 𝔻-BC-SGT, as a function of allophonic complexity and dimensionality.

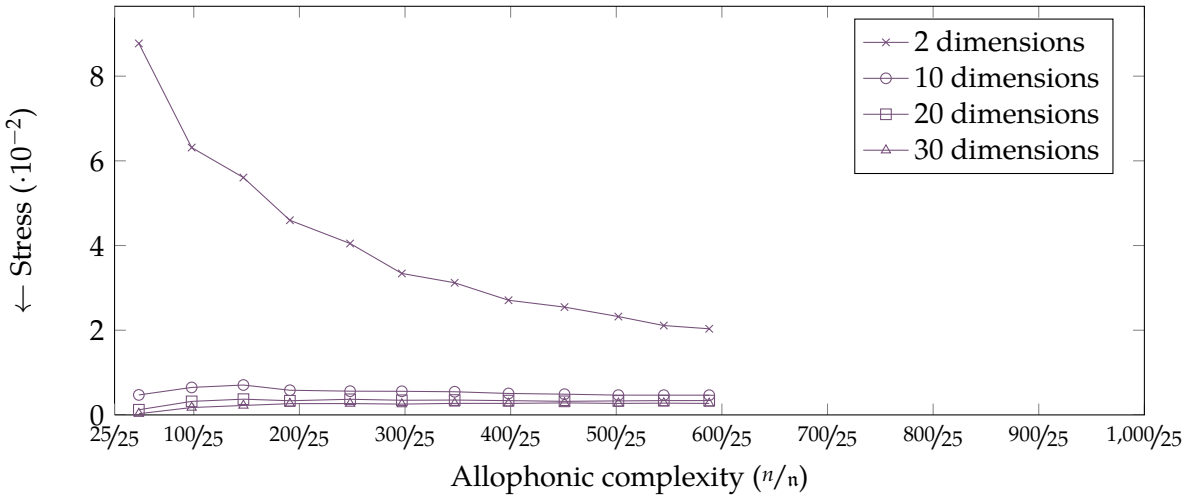


Figure 6.4 — Stress of the optimal metric configurations for 𝕃-WTP, as a function of allophonic complexity and dimensionality.

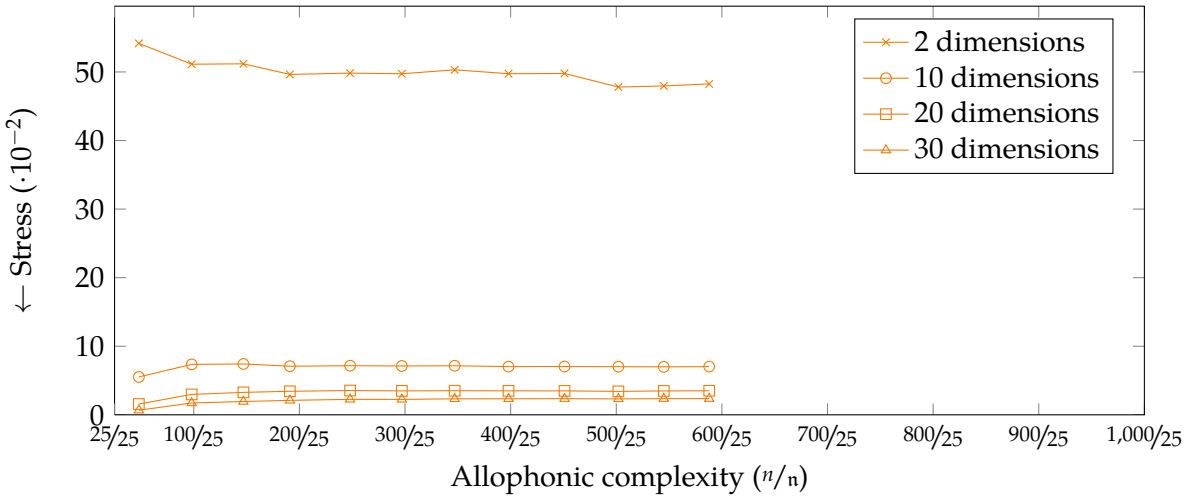


Figure 6.5 — Stress of the optimal metric configurations for the group model, as a function of allophonic complexity and dimensionality.

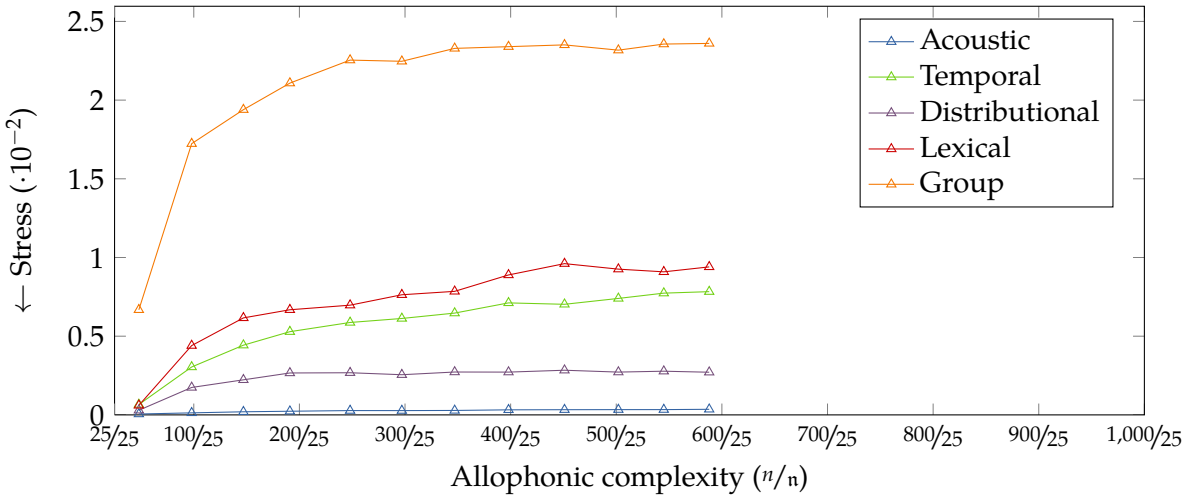


Figure 6.6 — Stress of the optimal metric configurations for the individual models and the group model in a 30-dimensional space, as a function of allophonic complexity.

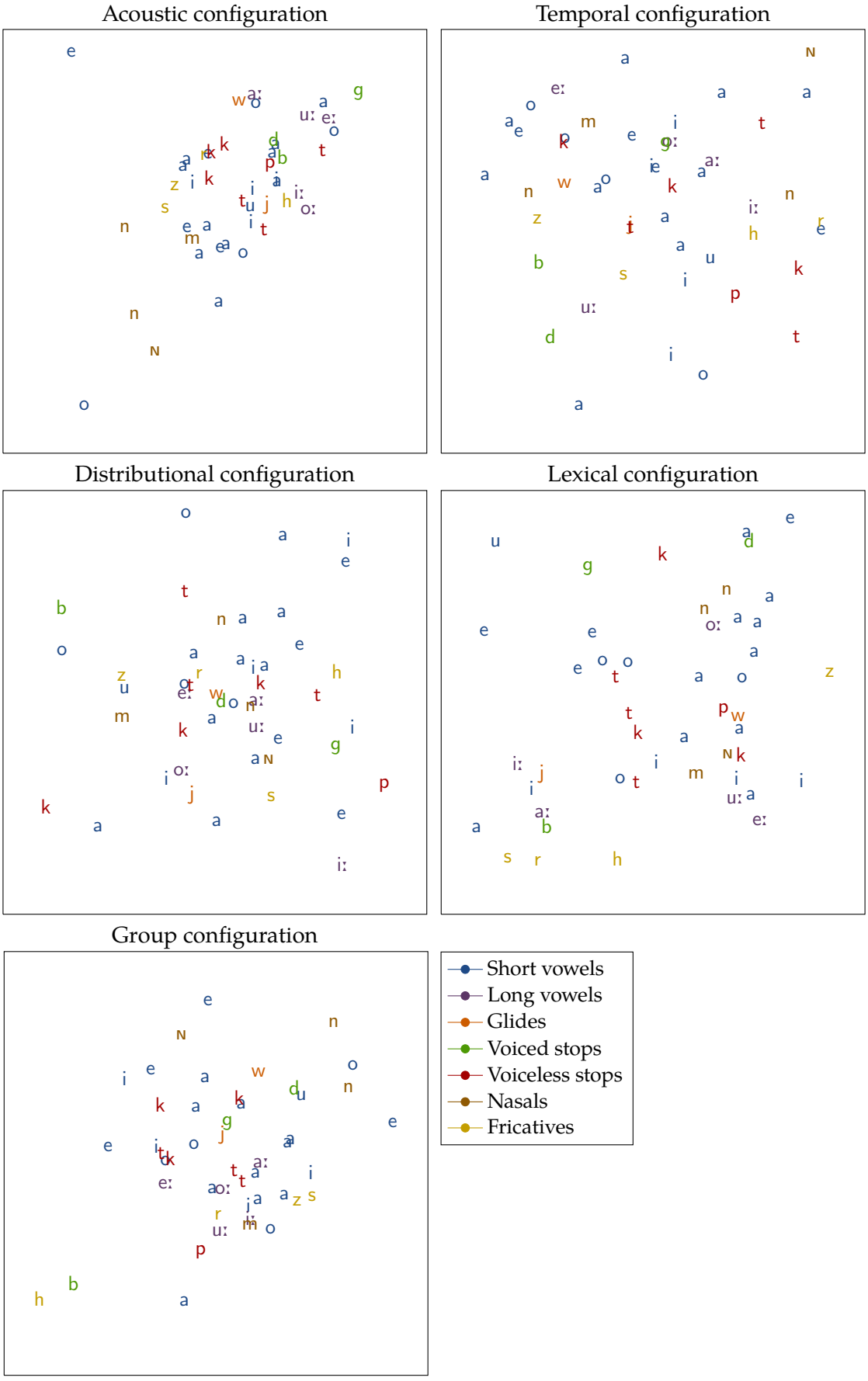


Figure 6.7 — Optimal 2-dimensional metric configurations at $^{48}/_{25}$ allophones per phoneme.

6.2 Prognoses of phonemehood

In this section, we address the question of whether there exists any structure in the data that resembles the phonemic partition. To this aim, we examine various cluster validity indices, i.e. quantitative criteria that assess the separation of the phonemic classes in the metric configurations. This section is thus the multiclass counterpart of the prognoses of allophony presented in Section 4.4.

Cluster validity indices Generally speaking, cluster validation refers to the quantitative evaluation of the quality of a clustering solution. Studies on cluster validation usually distinguish between three types of cluster validity indices (Halkidi et al., 2001; Aliguliyev, 2009): external criteria (whereby the solution is compared to a reference clustering), relative criteria (whereby the solution is compared to other solutions given by the same clustering technique but with different parameters), and internal criteria (whereby the solution is assessed against its own structural properties). In the interest of computing a prognosis of phonemehood, the clustering solution we want to assess is the reference phonemic partition $\mathfrak{P}$ and, hence, we focus in this section on internal cluster validity indices. A considerable number of such statistics have been proposed in the last decades and, in the words of Hubert & Arabie (1985; p. 193):

> ❝We will not try to review this literature comprehensively since that task would require the length of a monograph.❞

After reviewing various studies and textbooks that propose or compare clustering validity indices (Dunn, 1973; Hubert & Arabie, 1985; Milligan & Cooper, 1985; Rousseeuw, 1987; Bezdek & Pal, 1998; Jain et al., 1999; Weingessel et al., 1999; Strehl et al., 2000; Halkidi & Vazirgiannis, 2001; Halkidi et al., 2001; Tibshirani et al., 2001; Dudoit & Fridlyand, 2002; Chou et al., 2004; Kim & Ramakrishna, 2005; Mirkin, 2005; Legány et al., 2006; Rubinov et al., 2006; Meila, 2007; Jain, 2008; Manning et al., 2008; Aliguliyev, 2009; Barzily et al., 2009; Höppner, 2009; Reichart & Rappoport, 2009), it appears that the only internal cluster validity indices whose values were shown to lie in a finite range are the *normalized Hubert's statistic* (henceforth NHS; Hubert & Arabie, 1985) and the *overall average silhouette width* (henceforth OASW; Rousseeuw, 1987)—we did not consider unbounded indices on the grounds that, by definition, interpreting their values is more of an art than a science. However, the NHS yields very similar—if not identical—prognoses of phonemehood for all configurations and allophonic complexities. For this reason, we only report OASW-based prognoses in the present study, for the sake of brevity.

Silhouette widths The rationale behind Rousseeuw's (1987) concept of silhouettes is to assess the quality of a clustering solution when one is seeking compact and clearly separated clusters (cf. Duin, 1999; Pękalska et al., 2003) and when the point-to-point proximities are on a ratio scale—as in the case of Euclidean distance. It is worth noting that the computation of silhouette widths is only possible when the clustering solution comprises more than a single cluster—which is trivial in the case of phonemic inventories. The silhouette width of the phone $p_i \in P$ in the metric configuration $\mathbf{M}$ is given by

$$\text{silhouette}(p_i; \mathfrak{P}, \mathbf{M}) \equiv \frac{b(p_i; \mathfrak{P}, \mathbf{M}) - a(p_i; \mathfrak{P}, \mathbf{M})}{\max\{a(p_i; \mathfrak{P}, \mathbf{M}), b(p_i; \mathfrak{P}, \mathbf{M})\}} \tag{6.10}$$

where

$$a(p_i; \mathfrak{P}, \mathbf{M}) \equiv \sum_h \left[\frac{[\![p_i \in \mathfrak{p}_h]\!]}{|\mathfrak{p}_h| - 1} \sum_{p_j \in \mathfrak{p}_h} d_{ij}(\mathbf{M}) \right] \tag{6.11}$$

quantifies the average distance of p_i to the other allophones of the phoneme of which it is a realization, and

$$b(p_i; \mathfrak{P}, \mathbf{M}) \equiv \min_{\{\mathfrak{p}_h : p_i \notin \mathfrak{p}_h\}} \left[\frac{1}{|\mathfrak{p}_h|} \sum_{p_j \in \mathfrak{p}_h} d_{ij}(\mathbf{M}) \right] \tag{6.12}$$

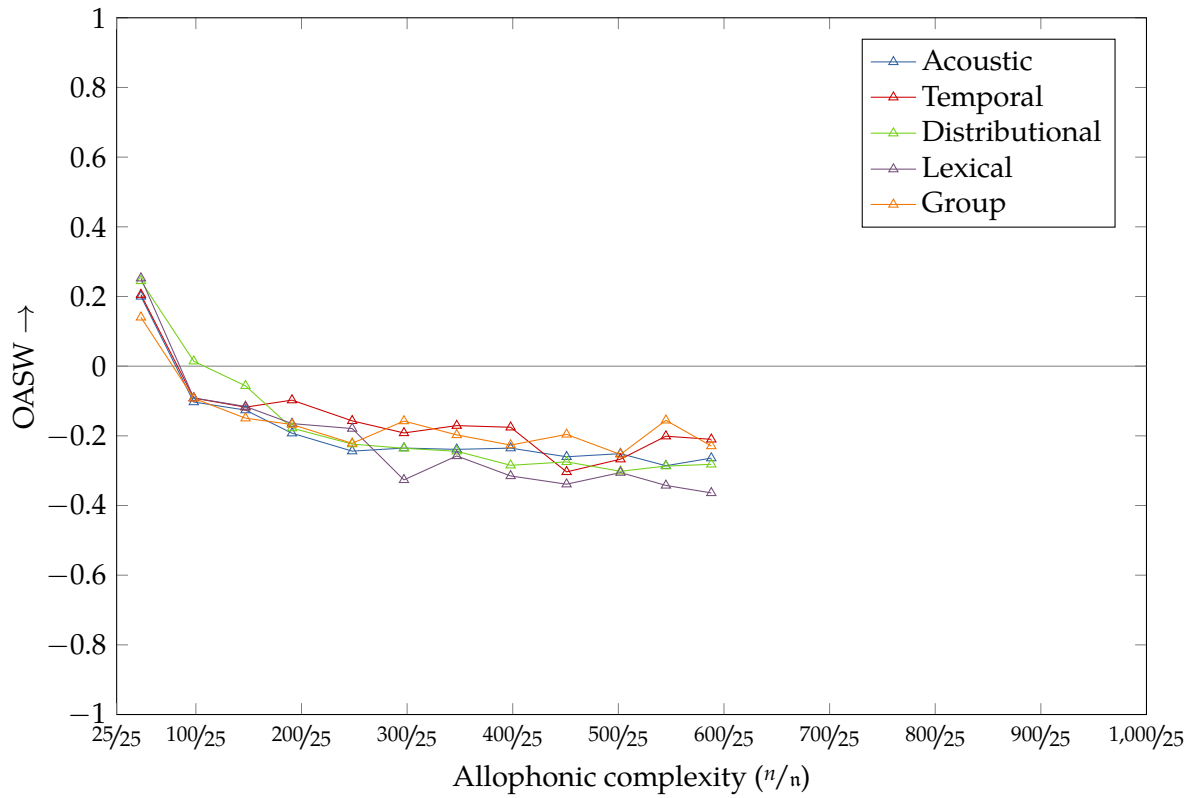


Figure 6.8 — Prognosis of phonemehood on the metric configurations, as a function of allophonic complexity.

is the minimum average dissimilarity of p_i to another phoneme than the one of which it is a realization. Silhouette widths lie in $[-1, 1]$ (Rousseeuw, 1987; p. 56). When silhouette$(p_i; \mathfrak{P}, \mathbf{M})$ is close to 1, there is little doubt that p_i has been assigned to a very appropriate cluster. When the silhouette width is close to -1, it appears that it was not natural to assign p_i to its cluster with respect to $\mathbf{M}$. When the silhouette width is about zero, there is no evidence that the point should have been assigned to any cluster.

Rousseeuw (1987) further defines the OASW of the whole clustering solution as the average of the silhouette widths of all points, i.e.

$$\text{OASW}(P, \mathfrak{P}; \mathbf{M}) \equiv \frac{\sum_i \text{silhouette}(p_i; \mathfrak{P}, \mathbf{\Delta})}{n}. \tag{6.13}$$

From the above definition, one can easily see that the OASW shares the same bounds as the individual silhouette widths. Accounting for instance weights, we propose the following straightforward extension

$$\text{OASW}(P, \mathfrak{P}; \mathbf{M}, \mathbf{w}) \equiv \frac{\sum_i w_i \, \text{silhouette}(p_i; \mathfrak{P}, \mathbf{M})}{\sum_i w_i} \tag{6.14}$$

It is worth noting that the computation of the OASW for the combined model amounts to replacing the individual configuration $\mathbf{M}$ with the group configuration $\mathbf{Z}$ in the equations.

Results The prognosis of phonemehood for the individual and the combined models are presented in Figure 6.8. The first observation is that this prognosis is not encouraging: all models' OASW is strictly negative from $^{147}/_{25}$ allophones per phoneme onwards. Although OASW values then appear to plateau as the allophonic complexity of the input further increases, negative values indicate that—on average—phonemes are neither compact nor clearly separated in the metric configurations. This prognosis of phonemehood corroborates the results of the preliminary $(\mathfrak{n}+1)$-ary classification experiments reported in Section 5.2: effective indicators of allophony do not appear to be effective indicators of phonemehood.

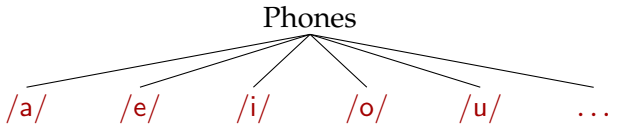


Figure 6.9 — Task diagram: $\mathfrak{n}$-ary phonemehood on phones. The task consists in predicting, for each phone in a given allophonic inventory, of which phoneme it is a realization.

It is worth noting that—for the first time in the present study—the acoustic indicator of allophony does not appear to be the indicator that better fits the target phonemic partition. Surprisingly, Figure 6.8 suggests that the group configuration may be the best—or, more accurately, the least bad—configuration for the discovery of phonemic clusters. This result is important in itself: though non-committal, OASW values show that combining indicators of allophony does not necessarily level down the information they contain.

6.3 Predicting phonemehood: $\mathfrak{n}$-ary classification task

In this section, our aim is to give empirical upper bounds on the learnability of phonemehood from the aforementioned Euclidean embeddings of the input allophonic inventories. To do so, we follow the methodology already presented in Sections 4.5 and 5.2, i.e. we rely on supervised learning and confirmatory overfitting. As illustrated in Figure 6.9, the task consists in predicting, for each phone in a given allophonic inventory, of which phoneme it is a realization.

6.3.1 Multinomial logistic regression

As discussed in Chapter 5, the major limitation of Peperkamp et al.'s (2006)'s framework lies in its pairwise formulation. In this chapter, we introduced a new representation for the input data whereby each phone $p_i \in P$ in a given allophonic inventory is represented as a point $\mathbf{m}_i \equiv (m_{i1}, \ldots, m_{iq})$ in a q-dimensional Euclidean space. Under this representation, phones are not only (indirectly) defined by their dissimilarity to other phones in the inventory, they are also materialized by points whose coordinates can be used as classification features or—in the terminology of logistic regression—predictor variables.

The model we use in these experiments is a classic multinomial logistic regression model, akin to the ones described in Section 5.2. Although we have already exposed the rationale and the assumptions of the multinomial logistic regression model in the previous chapter, we will hereby briefly revisit its formulation, adapting the notation to the data at hand.

Let $\{\pi_{i(1)}, \ldots, \pi_{i(\mathfrak{n})}\}$ denote the response probabilities associated to the $\mathfrak{n}$ phonemic classes $\mathfrak{P} \equiv \{\mathfrak{p}_1, \ldots, \mathfrak{p}_\mathfrak{n}\}$ for a given phone $p_i \in P$, where

$$\pi_{i(h)} \equiv \mathrm{P}(p_i \in \mathfrak{p}_h \mid \mathbf{m}_i) \tag{6.15}$$

denotes the probability of the phone p_i being a realization of the phoneme $\mathfrak{p}_h$ given its representation $\mathbf{m}_i \equiv (m_{i1}, \ldots, m_{iq})$ in a q-dimensional Euclidean space. The response probabilities satisfy

$$\sum_h \pi_{i(h)} = 1. \tag{6.16}$$

Using the q coordinates given by the optimal metric configuration $\mathbf{M}$ as the predictor variables, we have

$$\pi_{i(h)} = \frac{\exp(\alpha_h + \sum_l \beta_{hl} m_{il})}{\sum_{h'} \exp(\alpha_{h'} + \sum_l \beta_{h'l} m_{il})}. \tag{6.17}$$

As for the models presented in Section 5.2, the numerical algorithm we used to fit the parameters computes the $\mathfrak{n}(q+1)$ values that maximize the likelihood of the parameters given the configuration data (Agresti, 2007; Venables & Ripley, 2002; cf. `nnet::multinom`). Finally, the predicted phoneme of which a given phone p_i is a realization is merely defined as the most probable one.

6.3.2 Evaluation

In these experiments, as in the ones presented Sections 4.5 and 5.2, we are interested by the predictive power of the regression models. Cross-classifying the actual phonemic categories and the predicted categories yields a $\mathfrak{n} \times \mathfrak{n}$ contingency table $\mathbf{T} \equiv [t_{h'h}]$ where, by convention, h denotes the index over true classes and h' the index over predicted classes. Let $\hat{\mathfrak{P}} \equiv \{\hat{\mathfrak{p}}_1, \ldots, \hat{\mathfrak{p}}_\mathfrak{n}\}$ denote the predicted partition of the allophonic inventory. In each cell of the contingency table, the non-negative count $t_{h'h}$ is given by:

$$t_{h'h} \equiv \sum_i w_i [\![p_i \in \mathfrak{p}_h]\!] [\![p_i \in \hat{\mathfrak{p}}_{h'}]\!] \tag{6.18}$$

In order to assess the quality of a given model's predictions, we use quantitative criteria that were previously defined throughout the present study: the AIC (cf. Equation 4.45), the global accuracy (cf. Equation 5.16), pairwise precision (cf. Section 4.5.2), and pairwise recall (cf. Section 4.5.2). As hinted at in Section 5.2.3, multiclass predictions can indeed be considered as a series of pairwise decisions (Manning et al., 2008), i.e. one for each pair of phones. In that case, *TP* denotes the (weighted) number of allophonic pairs that were classified as realizations of the same phoneme, *TN* denotes the number of non-allophonic pairs that were classified as being realizations different phonemes, *FP* denotes the number of non-allophonic pairs that were classified as realizations of the same phoneme, and *FN* denotes the number of allophonic pairs that were classified as being realizations different phonemes.

Let $\mathbf{A} \equiv [a_{ij}]$ denote the predicted matrix of allophony derived from the predicted partition $\hat{\mathfrak{P}}$ of P. The predicted allophonic status a_{ij} of the pair of phones $\{p_i, p_j\} \in P$ is given by

$$a_{ij} \equiv [\![\exists\, h \in (1, 2, \ldots, \mathfrak{n}), \{p_i, p_j\} \subseteq \hat{\mathfrak{p}}_h]\!]. \tag{6.19}$$

The quantities *TP*, *TN*, *FP*, and *FN* can thus be computed from Equations 4.46, 4.49, 4.47, and 4.48. Consequently, the pairwise precision and the pairwise recall of a given model's predictions are given by Equations 4.51 and 4.52, respectively.

6.3.3 Results

The performance of the five multinomial logistic regression models—viz. acoustic, temporal, distributional, lexical, and combined—is presented in Figures 6.10, 6.11, 6.12, and 6.13 in terms of AIC, global accuracy, pairwise precision, and pairwise recall, respectively.

Goodness of fit Let us first consider the goodness of fit tests reported in Figure 6.10. As previously observed for binomial and flat-response pairwise multinomial logistic regression models (cf. Figures 4.21 and 5.4), the AIC of all models increases with the allophonic complexity of the input, meaning that the more phonemes have allophones, the less indicators of allophony (or an Euclidean embedding thereof) fit the phonemic partition described by $\mathfrak{P}$. Moreover, it appears that combining indicators of allophony is only effective from $^{248}/_{25}$ allophones per phoneme onwards—hence corroborating the prognosis of phonemehood presented in Figure 6.8.

These results confirm—to a certain extent—the relevance of using three-way MDS techniques for the combination of indicators of allophony: although it does not necessarily smear indicators' informativeness, it does not either necessarily improve the predictive power of the models.

Classification performance All three evaluation measures used to assess the quality of multinomial logistic regression models' performance on the $\mathfrak{n}$-ary classification task follow the same trend, as illustrated in Figures 6.11, 6.12, and 6.13. Whereas all five models yield very accurate predictions at the lowest allophonic complexity—viz. between 70% and 90% of correct predictions, depending on the model—their performance drops as the allophonic complexity of their input increases.

Unfortunately, the computational cost of SMACOF-based MDS limited our exploration of the learnability of phonemehood from metric configurations to $^{588}/_{25}$ allophones per phoneme.

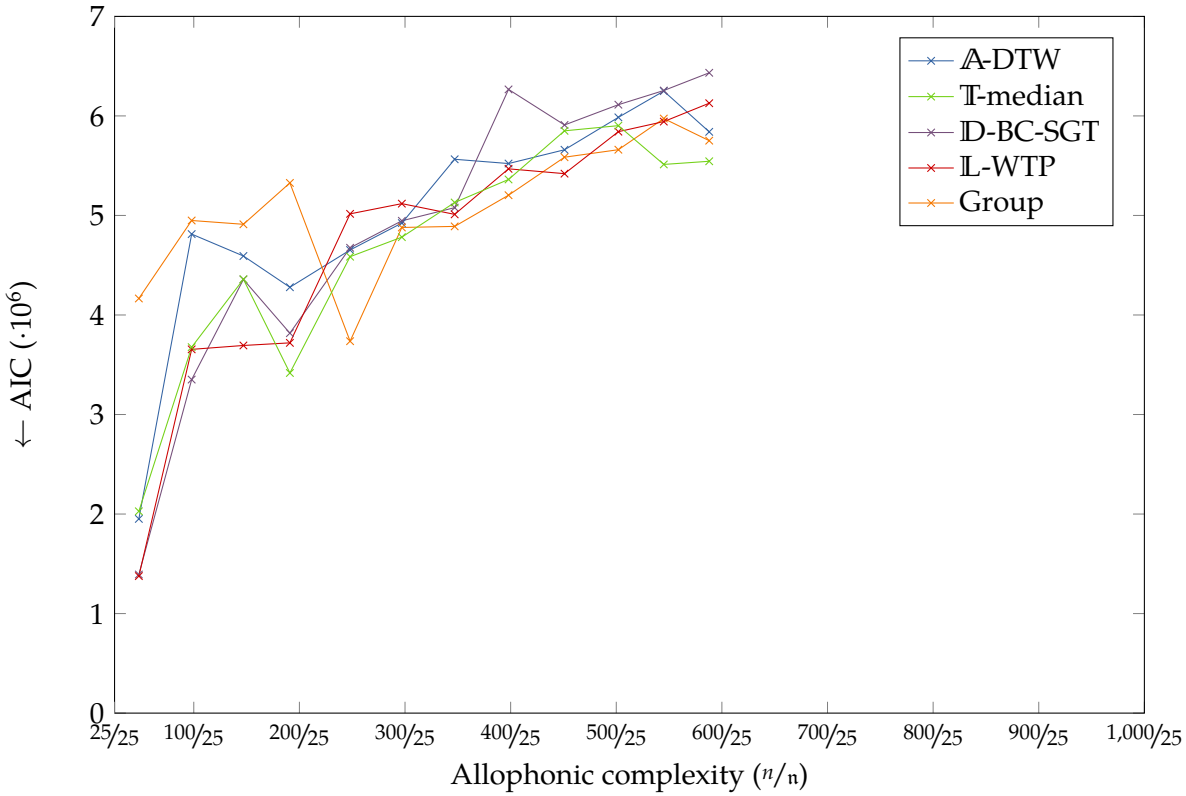


Figure 6.10 — AIC of the multinomial logistic regression models on the ɳ-ary classification task.

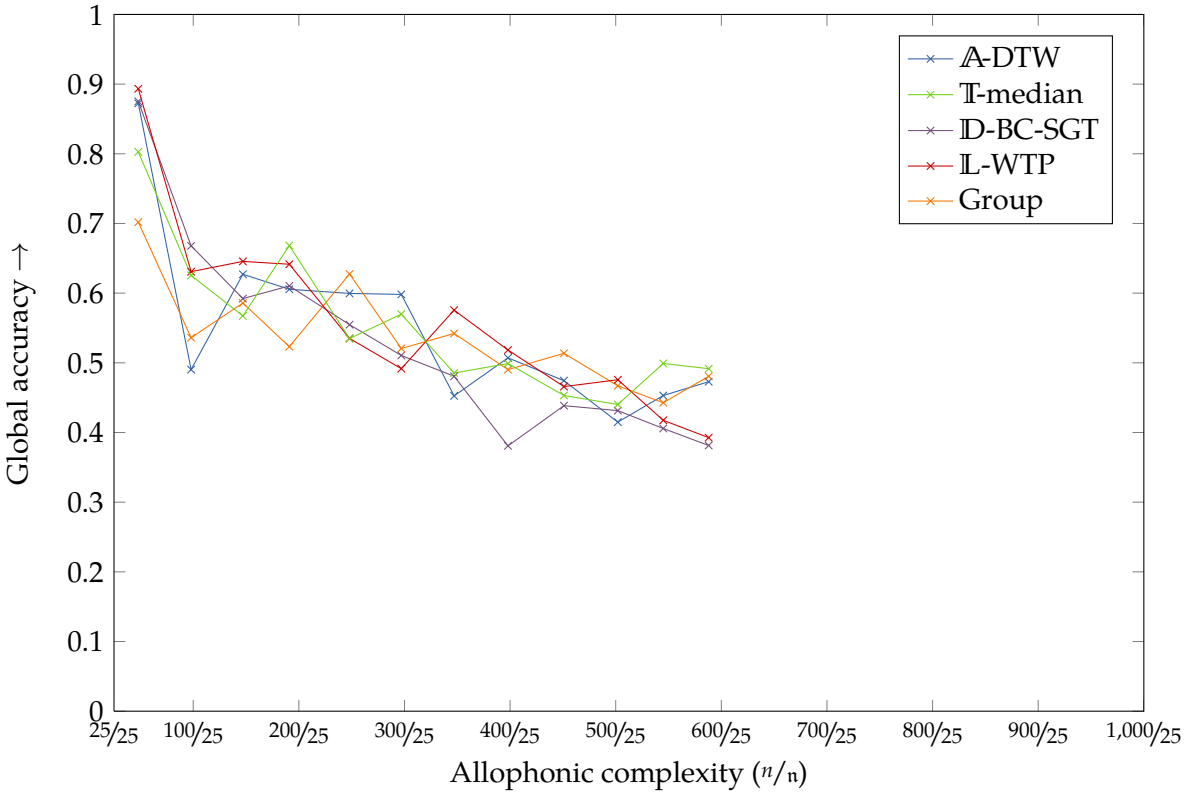


Figure 6.11 — Global accuracy of the multinomial logistic regression models on the ɳ-ary classification task.

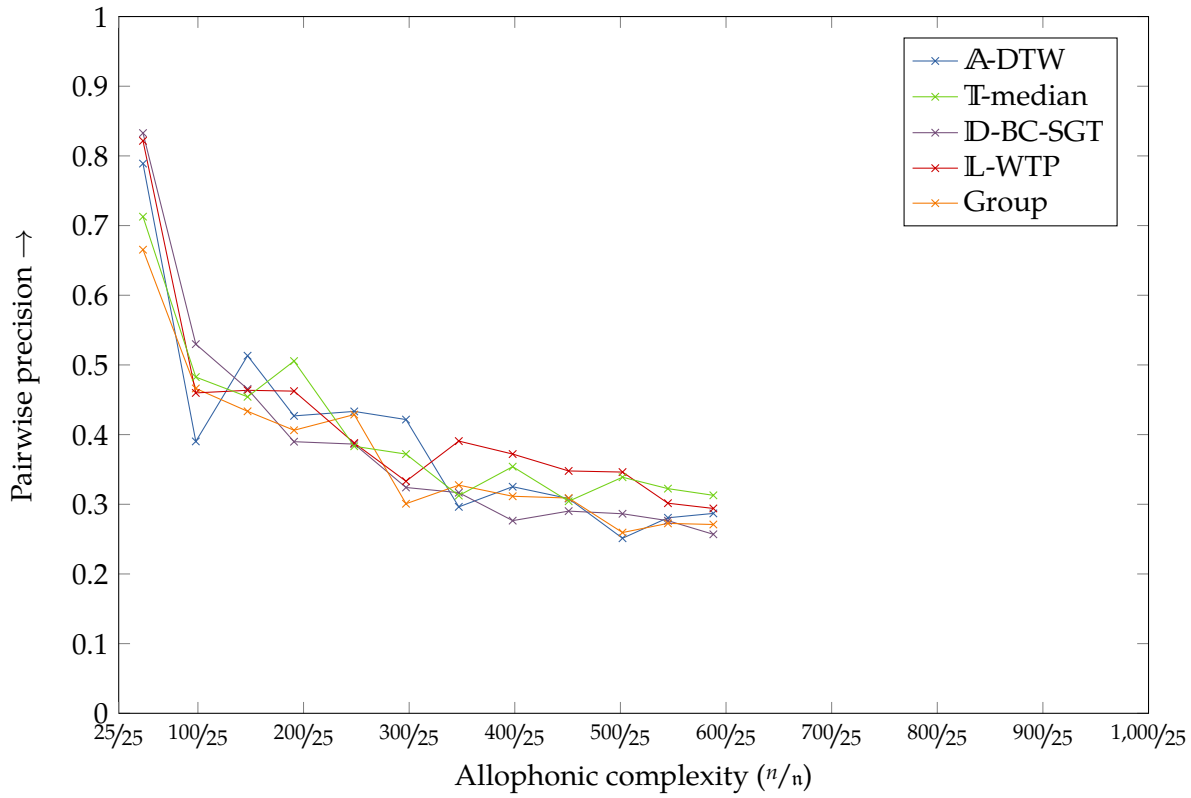


Figure 6.12 — Pairwise precision of the multinomial logistic regression models on the n-ary classification task.

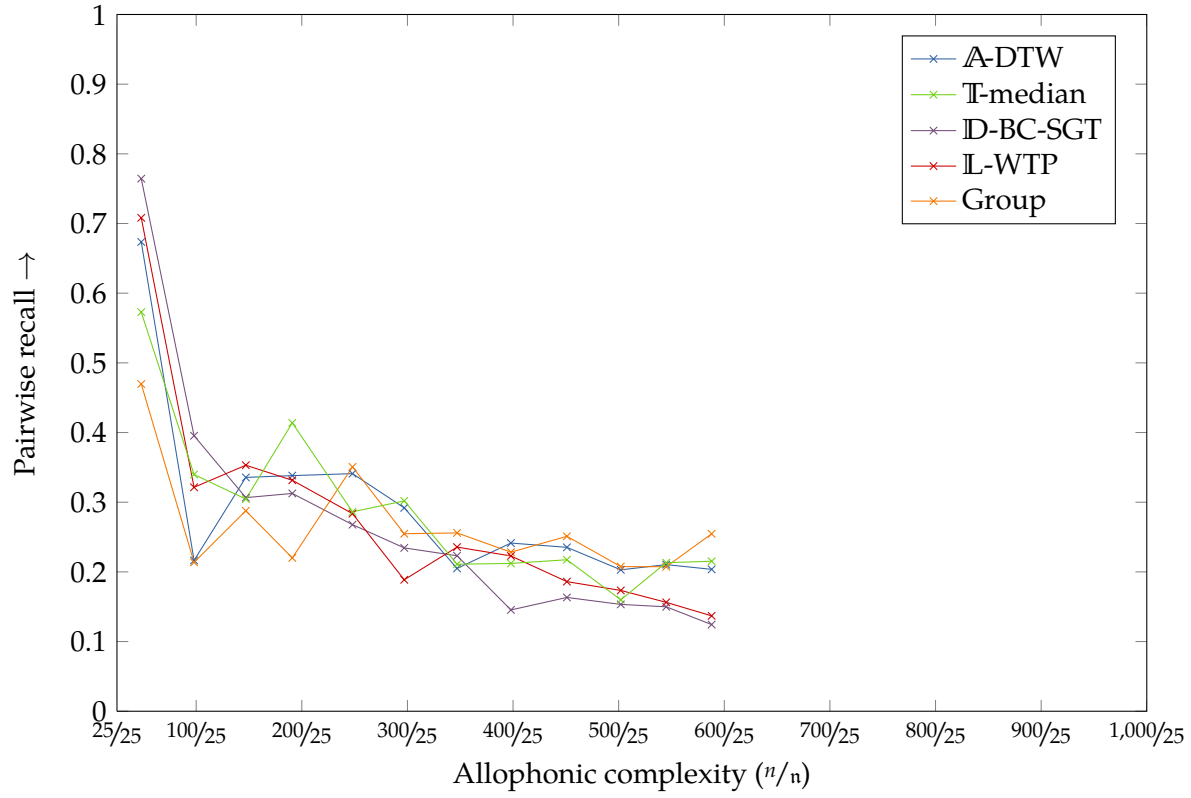


Figure 6.13 — Pairwise recall of the multinomial logistic regression models on the n-ary classification task.

Table 6.1 — Confusion table for the predictions of the multinomial logistic regression model on the group configuration, at $^{n}/_{\mathfrak{n}} = {}^{48}/_{25}$. Components may not sum to totals because of rounding.

↧	a	e	i	o	u	aː	eː	iː	oː	uː	w	j	b	d	g	p	t	k	m	n	ɴ	s	z	r	h
a	4.5	0.7	3.5	-	5.9	-	-	0.3	2.4	-	1.3	-	-	-	-	-	1.1	-	-	-	-	-	1.3	-	-
e	-	6.2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
i	3.7	-	3.8	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
o	-	-	-	8.9	-	-	-	-	-	-	-	-	0.8	-	-	-	-	-	-	-	-	-	-	-	-
u	-	-	-	-	0	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
aː	-	-	-	-	-	0	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
eː	-	-	-	-	-	-	1.5	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
iː	-	-	-	-	-	-	-	0	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
oː	-	-	-	-	-	-	-	-	0	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
uː	-	-	-	-	-	-	-	-	-	1.0	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
w	-	-	-	-	-	-	-	-	-	-	0	-	-	-	-	-	-	-	-	-	-	-	-	-	-
j	-	-	-	-	-	-	-	-	-	-	-	1.9	-	-	-	-	-	-	-	-	-	-	-	-	-
b	-	-	-	-	-	-	-	-	-	-	-	-	0	-	-	-	-	-	-	-	-	-	-	-	-
d	-	-	-	-	-	-	-	-	-	-	-	-	-	3.0	-	-	-	-	-	-	-	-	-	-	-
g	-	-	-	-	-	-	-	-	-	-	-	-	-	-	2.0	-	-	-	-	-	-	-	-	-	-
p	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0	-	-	-	-	-	-	-	-	-
t	0.6	-	-	1.4	-	-	-	-	-	-	-	-	-	-	-	-	7.5	-	-	-	-	-	-	-	-
k	3.0	-	-	-	-	0.4	-	-	-	-	-	-	-	-	-	-	-	6.6	-	-	-	-	-	-	-
m	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	3.4	-	-	-	-	-	-
n	1.4	-	1.7	-	-	-	-	-	-	-	-	-	-	-	-	0.4	-	-	-	5.7	-	-	-	-	-
ɴ	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	3.2	-	-	-	-
s	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	5.7	-	-	-
z	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0	-	-
r	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	3.7	-
h	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1.4

Table 6.2 — Confusion table for the predictions of the multinomial logistic regression model on the group configuration, at $^{n}/_{\mathfrak{n}} = {}^{588}/_{25}$. Components may not sum to totals because of rounding.

↧	a	e	i	o	u	aː	eː	iː	oː	uː	w	j	b	d	g	p	t	k	m	n	ɴ	s	z	r	h
a	3.9	0.7	2.4	2.6	-	-	ɛ	-	0.3	-	-	0.1	-	0.1	-	-	1.2	0.3	-	0.1	0.1	0.7	-	0.2	-
e	0.5	3.0	0.6	1.0	0.4	-	-	-	-	0.4	-	0.1	-	-	-	-	ɛ	0.4	0.3	0.4	0.3	0.4	-	0.2	0.2
i	1.0	0.9	3.7	1.1	-	-	0.6	-	-	-	-	0.4	-	-	-	-	0.8	0.6	-	ɛ	ɛ	0.3	-	0.3	-
o	3.7	0.3	0.6	3.0	-	-	-	-	0.4	-	-	0.2	0.2	0.2	-	-	0.9	0.4	-	0.1	0.1	0.7	-	-	-
u	0.1	-	0.4	0.3	5.2	-	-	-	-	-	-	-	-	-	-	-	0.3	0.4	0.2	-	ɛ	0.4	-	-	-
aː	-	-	-	-	-	0.4	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
eː	-	-	ɛ	-	-	-	0.8	-	-	-	-	-	-	-	-	-	-	ɛ	ɛ	0.2	0.2	-	-	-	-
iː	-	-	-	-	-	-	-	0	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
oː	0.1	-	-	ɛ	0.2	-	-	-	0.9	-	-	-	-	-	-	-	ɛ	-	-	-	0.1	ɛ	-	-	ɛ
uː	-	-	-	-	-	-	-	-	-	0.7	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
w	-	-	-	-	-	-	-	-	-	-	1.3	-	-	-	-	-	0.1	ɛ	-	-	-	ɛ	-	-	-
j	0.1	-	-	0.2	-	-	-	-	-	-	-	0.5	-	-	-	-	-	-	0.2	-	-	0.1	-	-	-
b	-	-	-	-	-	-	-	-	-	-	-	-	0.6	-	-	-	-	-	-	-	-	-	-	-	-
d	0.1	-	-	0.3	-	-	-	-	ɛ	-	-	-	-	1.8	-	-	-	ɛ	-	-	ɛ	ɛ	-	-	ɛ
g	-	-	-	-	-	-	-	-	-	-	-	-	-	-	2.0	-	-	-	-	-	-	-	-	-	-
p	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0	-	-	-	-	-	-	-	-	-
t	0.9	0.6	0.2	0.9	-	-	-	0.3	-	-	-	0.1	-	-	-	-	3.8	0.4	-	0.8	0.1	ɛ	-	0.8	-
k	0.7	0.4	0.1	0.3	-	ɛ	-	-	0.1	-	-	-	-	0.2	-	0.4	-	2.5	-	0.5	-	0.2	-	-	-
m	0.2	0.2	0.2	0.2	-	-	-	-	ɛ	-	-	-	-	-	-	-	0.5	-	1.7	-	0.1	ɛ	-	-	ɛ
n	0.8	0.4	0.2	ɛ	-	-	-	-	ɛ	-	-	-	-	0.4	-	-	0.1	0.6	-	3.6	0.4	ɛ	-	-	-
ɴ	0.1	-	0.4	0.1	-	-	-	-	-	-	-	-	-	-	-	-	0.2	0.4	-	-	1.5	-	-	-	-
s	1.0	0.3	0.2	ɛ	-	-	-	-	0.3	-	-	0.4	-	-	-	-	0.3	0.4	0.1	-	-	2.9	-	-	0.3
z	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.1	-	-	-	-	-	1.3	-	-
r	0.1	-	ɛ	0.4	-	-	-	-	0.3	-	-	-	-	0.3	-	-	0.1	0.1	0.4	ɛ	0.3	-	-	2.2	-
h	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.4	-	-	-	-	-	0.8

We are thus unable to assess whether the performance of our models keeps decreasing as the allophonic complexity of the input increases, or if it plateaus from $^{588}/_{25}$ allophones per phoneme onwards—as observed, for example, in Figures 5.7 and 5.7, for the $(\mathfrak{n}+1)$-ary classification tasks.

Finally, it is worth noting that Figures 6.11, 6.12, and 6.13 appear to be chaotic, as no definite pattern emerges regarding the identity of the optimal model, and the least performant model at a given allophonic complexity may well be the most performant one at the next complexity. We speculate that this is due to our (forced) choice to end the optimization process of each metric configuration after 100 SMACOF iterations—regardless of the concurrent gain in stress. Discussing the relative influence of SMACOF's convergence criteria is an issue we leave as a recommendation for future research.

Contingency tables For the sake of illustration, we report the contingency tables obtained for the combined model at minimal and maximal allophonic complexities in Tables 6.1 and 6.2, respectively. At $^{48}/_{25}$ allophones per phoneme, all allophones of 12 out of 25 phonemes (in green in the table header) are perfectly classified. By contrast, not a single allophone of 8 out of 25 phonemes was correctly classified (in red in the table header). Moreover, is is worth noting that misclassifications tend to consist in phones being predicted as being realizations of the most frequent phonemes in Japanese, viz. /a/, /i/, /o/, /t/, /k/, and /n/, as observed in Section 5.2.4. In other words, /a/ is the new $\otimes$. The figures reported in Table 6.2 confirm that this frequency-magnet effect still occurs at $^{588}/_{25}$ allophones per phoneme, as the same frequent phonemes account for most misclassifications.

All things considered, these results indicate that aggregating all phones in a given allophonic inventory into phoneme-like categories is not an easy task—even in a best-case scenario were labeled data is available (logistic regression is a supervised technique) and confirmatory overfitting is conducted. We can thus be but skeptical as to the learnability of phonemehood from the same data (and representation thereof), but in a unsupervised setup.

6.4 Predicting phonemehood: ?-ary clustering task

Up to this point, all experiments reported in this dissertation consisted in supervised learning techniques—actually variations on a single technique, i.e. logistic regression. However, from the perspective of early language acquisition, no labeled data should be available to the learner. The experiments we present in this section aim at assessing whether compact and well-separated clusters can be inferred from the indicators' metric configurations. As emphasized by Höppner (2009; p. 386):

> ❝We are *not* interested in artificially dividing the [objects] into similar groups, but we want to know if the data itself supports a partition into different groups,❞

hence the jocular arity in the title of this section.

Definitions and objectives Clustering—a classic approach to unsupervised learning—refers to the problem of trying to find hidden structure in unlabeled data (e.g. Jain et al., 1999; Mirkin, 2005). In the case at hand, we aim at partitioning the allophonic inventory P into a set $\Omega \equiv \{\omega_1, \ldots, \omega_\eta\}$ of η collectively exhaustive and mutually exclusive clusters. As previously discussed throughout the present study, the optimal partition of any given allophonic inventory P is the phonemic inventory $\mathfrak{P}$ of the language at hand. Moreover, the quality of the clustering solution Ω is bounded by the following limiting cases (Reichart & Rappoport, 2009): the single cluster solution, whereby all phones are assigned to the same unique cluster, i.e.

$$\Omega = \{P\}, \tag{6.20}$$

and the singletons solution, whereby each phone is assigned to a cluster of its own, i.e.

$$\Omega = \{\{p_i\} : p_i \in P\}. \tag{6.21}$$

Furthermore, we refer to a partition whereby each cluster contains only allophones as being *homogeneous*, and to a partition whereby, for each phoneme, all its allophones are assigned to the same cluster as being *complete* (Meila, 2007; Rosenberg & Hirschberg, 2007; Reichart & Rappoport, 2009). Consequently, whereas the single cluster solution is only complete and the singletons solution is only homogeneous, the reference phonemic partition $\mathfrak{P}$ is both fully homogeneous and fully complete.

6.4.1 Density-based clustering with DBSCAN

In order to assess whether the data at hand supports a partition into different groups, we use the DBSCAN algorithm (Ester et al., 1996), a classic density-based clustering algorithm that is able to discover clusters of arbitrary shapes. Whereas other density-based clustering algorithms have been proposed (e.g. CLARANS; Ng & Han, 1994), DBSCAN's main attraction arises from its efficiency on large datasets, as its average time complexity is in $\mathcal{O}(n \log_2 n)$.

Another remarkable property of DBSCAN is that this algorithm does not require locations estimates, but only object-to-object dissimilarities. In the case at hand, this property allows for the comparison of clustering solutions obtained for both the original dissimilarities and the computed Euclidean distances. In this section, let $\Delta \equiv [\delta_{ij}]$ denote the input dissimilarity or distance matrix, regardless of its true nature. The original formulation of DBSCAN does not accommodate weighted objects. For this reason, we hereby give a detailed presentation of the weighted extension of DBSCAN we propose, emphasizing the discrepancies with Ester et al.'s (1996) formulation.

Density-based clustering The notion of spatial density in DBSCAN resembles the notion of population density in demographic studies, in the sense that it is defined as a combination of population and volume: given objects must be close enough and numerous enough in order to form a dense cluster. In DBSCAN, density is further defined by two parameters, viz. $\epsilon \in \mathbb{R}^+$ which controls how far two points can be so that they may still be assigned to the same cluster, and $\phi \in \mathbb{N}$ which controls how many close-enough points are required to start a cluster.

One of the two major concepts shared by DBSCAN and its extensions is the ϵ-neighborhood of an object. Using Kailing et al.'s (2004b) notation, let $\mathcal{N}_\epsilon^{\mathbf{\Delta}}(p_i)$ denote the ϵ-neighborhood of a given phone $p_i \in P$ with respect to the representation $\mathbf{\Delta}$. The ϵ-neighborhood of p_i is defined as the set of objects whose distance to p_i is less than or equal to ϵ, i.e.

$$\mathcal{N}_\epsilon^{\mathbf{\Delta}}(p_i) \equiv \left\{ p_j \in P : \delta_{ij} \leq \epsilon \right\}. \tag{6.22}$$

It is worth highlighting that, according to this definition, any object belongs to its own ϵ-neighborhood.

The other major concept allowing for the definition of density in DBSCAN is core objects. In the original formulation of the algorithm, a core object with respect to $\mathbf{\Delta}$, ϵ, and ϕ was defined as an object whose ϵ-neighborhood contains at least ϕ other phones, that is

$$\text{core}(p_i; \mathbf{\Delta}, \epsilon, \phi) \equiv [\![\phi \leq |\mathcal{N}_\epsilon^{\mathbf{\Delta}}(p_i)|]\!]. \tag{6.23}$$

In the case at hand, we amend this definition so that each phone's instance weight impacts the computation of the density in its vicinity. One solution would consist in replicating w_i times the point $\mathbf{m}_i$ representing a given phone $p_i \in P$. However, this solution is not desirable at the time complexity of the clustering procedure would thus be in $\mathcal{O}(\sum_i w_i \log_2 \sum_i w_i)$—rather than in $\mathcal{O}(n \log_2 n)$. In our weighted extension of DBSCAN, a core object with respect to $\mathbf{\Delta}$, $\mathbf{w}$, ϵ, and ϕ is therefore defined as an object whose ϵ-neighborhood weighs at least ϕ, i.e.

$$\text{core}(p_i; \mathbf{\Delta}, \mathbf{w}, \epsilon, \phi) \equiv \left[\!\!\left[\phi \leq \sum_{p_j \in \mathcal{N}_\epsilon^{\mathbf{\Delta}}(p_i)} w_j \right]\!\!\right] \tag{6.24}$$

It is worth noting that the latter definition is similar to the weight function used in Sander et al.'s (1998) density-based clustering algorithm.

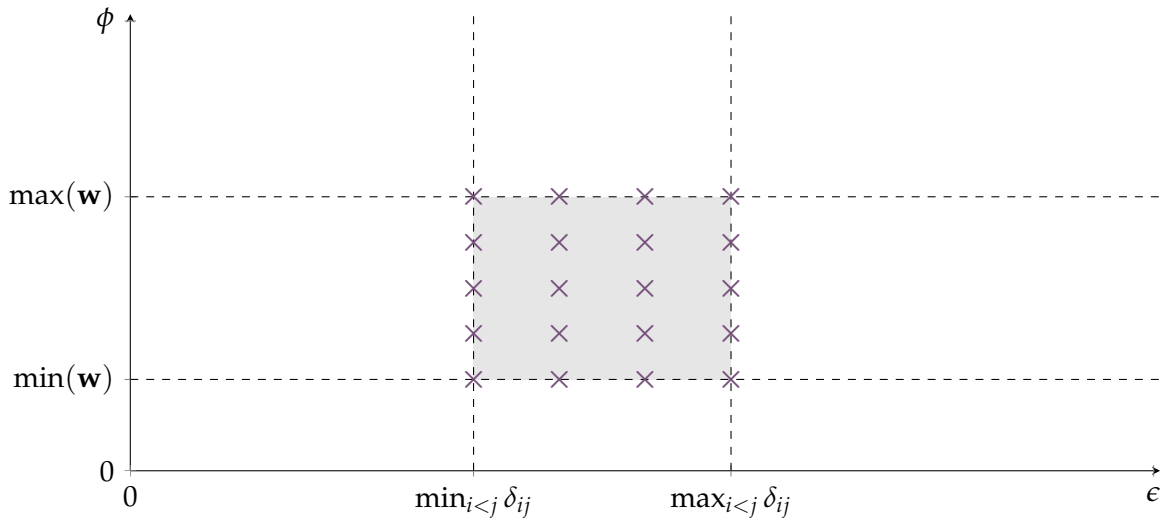


Figure 6.14 — Schematic diagram of the grid search method. The gray area marks the subspace delimited by the theoretical bounds we defined. The plum marks denote linearly spaced trials.

The following definitions, true to the ones given by Ester et al. (1996), build upon the concept of ϵ-neighborhood and core object to define density-based clusters. First, a phone p_j is directly density-reachable from a phone p_i with respect to $\boldsymbol{\Delta}$, $\mathbf{w}$, ϵ, and ϕ if p_i is a core object with respect to $\boldsymbol{\Delta}$, $\mathbf{w}$, ϵ, and ϕ and if p_j belongs to the ϵ-neighborhood of p_i with respect to $\boldsymbol{\Delta}$. Furthermore, a phone p_j is density-reachable from a phone p_i with respect to $\boldsymbol{\Delta}$, $\mathbf{w}$, ϵ and ϕ if there is a chain of phones $p_j, \ldots, p_i$ such that each phone is directly density-reachable from the preceding one with respect to $\boldsymbol{\Delta}$, $\mathbf{w}$, ϵ, and ϕ. Then, two phones p_i and p_j are density-connected with respect to $\boldsymbol{\Delta}$, $\mathbf{w}$, ϵ and ϕ if there is a phone $p_{i'}$ such that both p_i and p_j are density-reachable from $p_{i'}$ with respect to $\boldsymbol{\Delta}$, $\mathbf{w}$, ϵ and ϕ. Finally, a cluster is defined as a set of density-connected phones which is maximal with respect to density-reachability. In other words, a density-based cluster is defined as a set of phones whose ϵ-neighborhood intersect—or are tangent to each other—and whose global weight is greater than or equal to ϕ.

According to the original formulation of DBSCAN, the subset of phones in P that do not belong to any density-based cluster would be defined as noise. However, as the data at hand contains no true noise but—at worst—isolated, non-core phones, we assign each of these phones to a cluster of their own.

Model-fitting Before turning to the evaluation of the clustering solutions discovered by DBSCAN, we need to define how the values of the parameters ϵ and ϕ were set in our experiments. Whereas various heuristics have been proposed to estimate these parameters (e.g. Ester et al., 1996; Esmaelnejad et al., 2010), none of these procedures is fully automatic—to say nothing of the ones that propose to replace ϵ or ϕ with yet another parameter. Moreover, because no density-based clustering experiment has yet been reported on indicators of allophony, we are interested in assessing how the performance of the algorithm varies as a function of the parameter values.

Because the true search space for each parameter is infinite, we can but use a strategy to test a finite and reasonably-sized subset of parameter combinations. Manual search (i.e. chosing some values by ourselves) is impractical as we have virtually no a priori knowledge of how the clusters are spatially organized in the metric configuration of a given indicator (or combination thereof), and how such structures evolve as the allophonic complexity of the input increases. Random search has been shown to be both efficient and effective (Bergstra & Bengio, 2012) but, by definition, it gives a low degree of insight into the full range of possible combinations. For these reasons, we performed a bounded and discretized grid search on the parameter values. Grid searching refers to exhaustively searching through a finite subset of the parameter space

(e.g. Hsu et al., 2010), as illustrated in Figure 6.14. It is reliable in low-dimensional parameter spaces—as in the case at hand, where we have 2 parameters—and, from a technical perspective, it is simple to implement and parallelization is trivial (Bergstra & Bengio, 2012).

It order to guarantee the tractability and the relevance of our experiments, we first define, for each parameter ϵ and ϕ, a range of relevant values to be tested. To our knowledge, the following a priori bounds have never been described, neither for DBSCAN nor for derivatives thereof (Ester et al., 1996; Sander et al., 1998; Ankerst et al., 1999; Wang & Hamilton, 2003; Wang et al., 2004; Kailing et al., 2004a,b; Gorawski & Malczok, 2006a,b; Ruiz et al., 2007)—it is worth noting, however, that Esmaelnejad et al.'s (2010) heuristic for setting the value of ϵ has a similar rationale.

As aforestated, $\epsilon \in \mathbb{R}^+$ controls how far two phones can be so that they may still be assigned to the same cluster. Thus, it mostly interacts with the distances or dissimilarities in $\boldsymbol{\Delta}$. If $\epsilon < \min_{i<j} \delta_{ij}$, no two phones are ϵ-reachable and, hence, each phone's ϵ-neighborhood is reduced to the phone itself and the clustering solution can not be better than the singletons solution. If $\epsilon = \max_{i<j} \delta_{ij}$, then the farthest phones are ϵ-reachable. Thus, all other things being equal, clustering with $\epsilon > \max_{i<j} \delta_{ij}$ can not yield better solutions than the one obtained with $\epsilon = \max_{i<j} \delta_{ij}$. The relevant range of values for ϵ is hence bounded by the attested range of distance or dissimilarity values in $\boldsymbol{\Delta}$:

$$\min_{i<j} \delta_{ij} \leq \epsilon \leq \max_{i<j} \delta_{ij}. \tag{6.25}$$

On the other hand, $\phi \in \mathbb{N}$ controls how many close-enough phones are required to start a cluster. Thus, it mostly interacts with the weights in $\mathbf{w}$ to select core objects. The minimum weight of all phones' ϵ-neighborhoods is $\min(\mathbf{w})$. Therefore, all other things being equal, clustering with $\phi < \min(\mathbf{w})$ can not yield clustering solutions different than the one obtained with $\phi = \min(\mathbf{w})$. Similarly, if $\phi > \max(\mathbf{w})$, then no phone can be a core object and no clustering solution can be different from the singletons solutions. The relevant range of values for ϕ is hence bounded by the attested range of distance or dissimilarity values in $\mathbf{w}$:

$$\min(\mathbf{w}) \leq \phi \leq \max(\mathbf{w}). \tag{6.26}$$

Relying on these a priori bounds on the parameters ϵ and ϕ, we tested 100 linearly-spaced values for each parameter, i.e. we performed a grid search over $100^2 = 10,000$ DBSCAN trials for each indicator of allophony (or combination thereof), and for each allophonic complexity.

6.4.2 Evaluation

Because providing a thorough examination of thousands of clustering solutions is unfeasible—if relevant at all—we only report the performance of DBSCAN for the optimal clustering solution computed for each indicator and allophonic complexity.

Contingency table Cross-classifying a given clustering solution Ω and the reference phonemic partition $\mathfrak{P}$ yields an $\eta \times \mathfrak{n}$ contingency table $\mathbf{T} \equiv [t_{oh}]$ where $o \in (1, 2, \ldots, \eta)$ denotes the index over the predicted clusters, and $h \in (1, 2, \ldots, \mathfrak{n})$ denotes the index over the phonemes of the target language. In each cell, the non-negative count t_{oh} is given by the (weighted) number of allophones of $\mathfrak{p}_h$ that were assigned to the cluster ω_o, i.e.

$$t_{oh} \equiv \sum_i w_i [\![p_i \in \mathfrak{p}_h]\!] [\![p_i \in \omega_o]\!]. \tag{6.27}$$

Objective function The quantitative criterion we use as the objective function to retrieve the optimal clustering solution in the grid search is known as the *normalized variation of information* (henceforth NVI; Reichart & Rappoport, 2009). NVI is an information-theoretic measure given by:

$$\mathrm{NVI}(\Omega; \mathfrak{P}) \equiv \begin{cases} \dfrac{\mathrm{H}(\mathfrak{P} \mid \Omega) + \mathrm{H}(\Omega \mid \mathfrak{P})}{\mathrm{H}(\mathfrak{P})} & \text{if } \mathrm{H}(\mathfrak{P}) \neq 0 \\ \mathrm{H}(\Omega) & \text{otherwise} \end{cases} \tag{6.28}$$

where

$$\mathrm{H}(\mathfrak{P}) \equiv -\sum_{h} \left[\frac{\sum_o t_{oh}}{\sum_i w_i} \log_2 \frac{\sum_o t_{oh}}{\sum_i w_i} \right], \tag{6.29}$$

$$\mathrm{H}(\mathfrak{P} \mid \Omega) \equiv -\sum_{o} \sum_{h} \left[\frac{t_{oh}}{\sum_i w_i} \log_2 \frac{t_{oh}}{\sum'_h t_{oh}} \right], \tag{6.30}$$

$$\mathrm{H}(\Omega) \equiv -\sum_{o} \left[\frac{\sum_h t_{oh}}{n} \log_2 \frac{\sum_h t_{oh}}{\sum_i w_i} \right], \tag{6.31}$$

$$\mathrm{H}(\Omega \mid \mathfrak{P}) \equiv -\sum_{o} \sum_{h} \left[\frac{t_{oh}}{\sum_i w_i} \log_2 \frac{t_{oh}}{\sum'_o t_{oh}} \right]. \tag{6.32}$$

Although discussing the rationale of the computation of NVI is beyond the scope of this dissertation, it is worth highlighting that NVI values decrease as the clustering solution becomes more complete and more homogeneous and, hence, that the reference phonemic partition $\mathfrak{P}$ has a NVI equal to 0. For each indicator of allophony (or combination thereof) and allophonic complexity, the optimal clustering solution is thus the one that minimizes the NVI. Moreover, the single cluster solution has a NVI equal to 1. Although Reichart & Rappoport (2009) give no upper bound for NVI, they indicate that all clustering solutions of higher quality than the single cluster solution are scored by NVI in the range $[0, 1]$.

Because NVI is an external cluster validity index, one could object that performing a grid search whose objective function relies on the reference clustering solution—i.e. $\mathfrak{P}$—is akin to performing supervised learning. It is however worth noting that each DBSCAN trial is truly unsupervised. Furthermore, in light of the results presented in Sections 6.2 and 6.3, we know that none of the metric configurations we computed from our indicators of allophony supports a complete and homogeneous phonemic partition of the allophonic inventories. The results to be further reported in this section are thus the best among the worst.

Purity For the sake of comparability with previously reported experiments in this study, we also use other evaluation measures to assess the quality of the density-based clustering solutions. Whereas accuracy is only relevant in supervised learning setups—i.e. when each predicted category can be matched to a reference category—*purity* can be interpreted as its unsupervised counterpart (Manning et al., 2008). To compute the purity of a clustering solution, each cluster is assigned to the phoneme which is most frequent in the cluster, and then the accuracy of this assignment is measured by counting the (weighted) number of correctly assigned phones and dividing by the number of phones. Formally, we have:

$$\mathrm{purity}(\mathbf{T}) \equiv \frac{1}{\sum_i w_i} \sum_{o} \left(\max_{h} t_{oh} \right) \tag{6.33}$$

From this definition, one can easily see that purity values lie in $[0, 1]$: bad clustering solutions have purity values close to 0, and the reference phonemic partition $\mathfrak{P}$ has a purity of 1. As emphasized by (Reichart & Rappoport, 2009), purity evaluates homogeneity, but does not evaluate completeness at all. Whereas the purity of the singletons solutions is, by definition, equal to 1, the purity of the single cluster solution depends on the dataset at hand. More precisely, it is given by the ratio of the frequency of the most frequent phoneme to the sum of all phonemes' frequency, i.e.

$$\frac{\max_h \sum_i w_i [\![p_i \in \mathfrak{p}_h]\!]}{\sum_i w_i} = \frac{212554}{1591966} \approx 0.13 \tag{6.34}$$

Pairwise evaluation As with any multiclass problem, the predictions can be considered as a series of pairwise decision. Using the same rationale as in Section 6.3, the quantities *TP*, *TN*, *FP*,

and *FN* are given by

$$TP \equiv \sum_{h,o} \binom{t_{ho}}{2} \tag{6.35}$$

$$FP \equiv -TP + \sum_{h} \binom{\sum_o t_{ho}}{2} \tag{6.36}$$

$$FN \equiv -TP + \sum_{o} \binom{\sum_h t_{ho}}{2} \tag{6.37}$$

$$TN \equiv \binom{\sum_{h,o} t_{ho}}{2} - TP - FP - FN \tag{6.38}$$

Consequently, the pairwise precision and the pairwise recall of a given model's predictions are given by Equations 4.51 and 4.52, respectively.

6.4.3 Results

The performance of the density-based clustering model on all aforementioned metric configurations is presented in Figures 6.16, 6.17, 6.18, and 6.19 in terms of NVI, purity, precision, and recall, respectively.

Grid search Let us first consider the results of the grid searches, as illustrated in Figure 6.15 where—for the sake of brevity—we only present the surface responses for a subset of the tested allophonic complexities. For each model and allophonic complexity, the surface response is a graphical representation of the result of the corresponding grid search whereby the observed NVI values for all tested parameter values are cross-classified. NVI values are mapped to a linear grayscale whereby values equal to 0 are mapped to black and values greater than or equal to 1 to white. Because NVI is to be minimized, dark areas thus indicate effective parameter combinations.

All surface response call for a general observation: the value of ϵ is irrelevant for the purpose of clustering any given allophonic inventory into a phoneme-like partition. Indeed, if it is not completely blank, each response surface shows a horizontal pattern, meaning that only the parameter on the y-axis (i.e. ϕ) has an impact on the quality of the clustering solution. Because ϕ mostly interacts with the weights in $\mathbf{w}$, this result corroborates—once again—the influence of the frequency-based instance weights on the discovery of phonemes.

Another interesting observation is that, except for the acoustic indicator, the 30-dimensional metric configurations appear to yield better density-based clustering solutions than the corresponding original dissimilarities. Whereas a dark horizontal pattern appears, for instance, in the response surfaces for the metric configurations of the distributional and lexical indicators, the corresponding response surfaces for their original dissimilarities are completely blank—regardless of the allophonic complexity of their input. By contrast, Figure 6.15 suggests that the effectiveness of the acoustic indicator for the discovery of phonemes suffers—to some extent—from the Euclidean embedding, as the corresponding response surfaces fade to white as allophonic complexity increases. This result is important in itself, and we consider it to be a validation of our MDS-based reformulation of Peperkamp et al.'s (2006) framework. Indeed, the response surfaces presented in Figure 6.15 show that embedding indicators of allophony in a multidimensional Euclidean space does not smear the information they contain—to say nothing of the availability of three-way extension allowing for the combinations of various indicators.

Clustering quality All other evaluation measures give a similar assessment of the quality of the density-based phoneme-like clustering solutions, as illustrated in Figures 6.16, 6.17, 6.18, and 6.19. Except for the acoustic indicator, DBSCAN always outputs the singletons solution, i.e. the limiting case whereby each phone in the allophonic inventory is assigned to its own cluster.

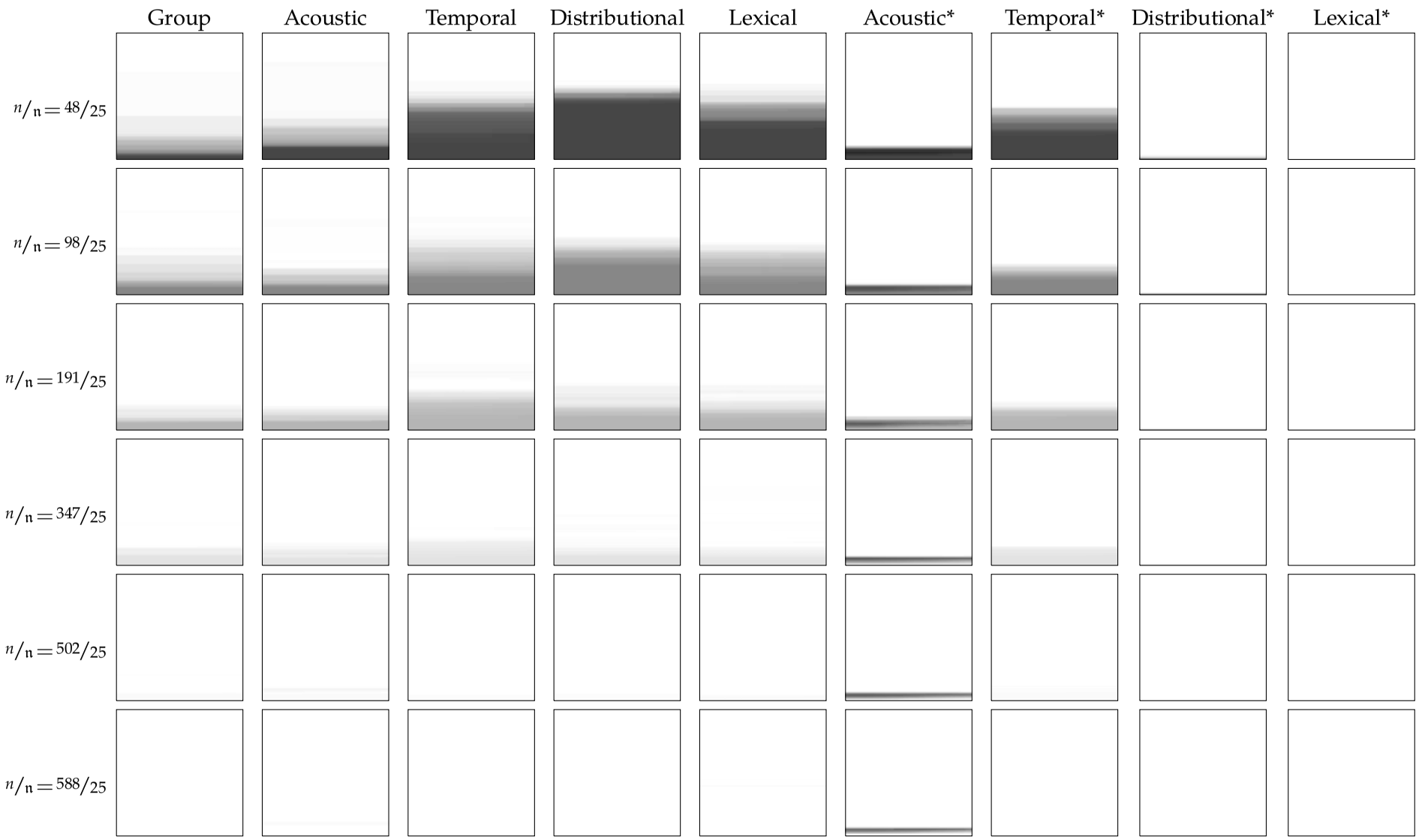


Figure 6.15 — Response surface of the grid searches for DBSCAN's optimal parameters, as a function of allophonic complexity. For each response surface, ϵ is on the x-axis and ϕ on the y-axis. NVI values are mapped to a linear grayscale whereby values equal to 0 are mapped to black and values greater than or equal to 1 to white. Starred column headers indicate that the raw dissimilarities where used, rather than the metric configurations.

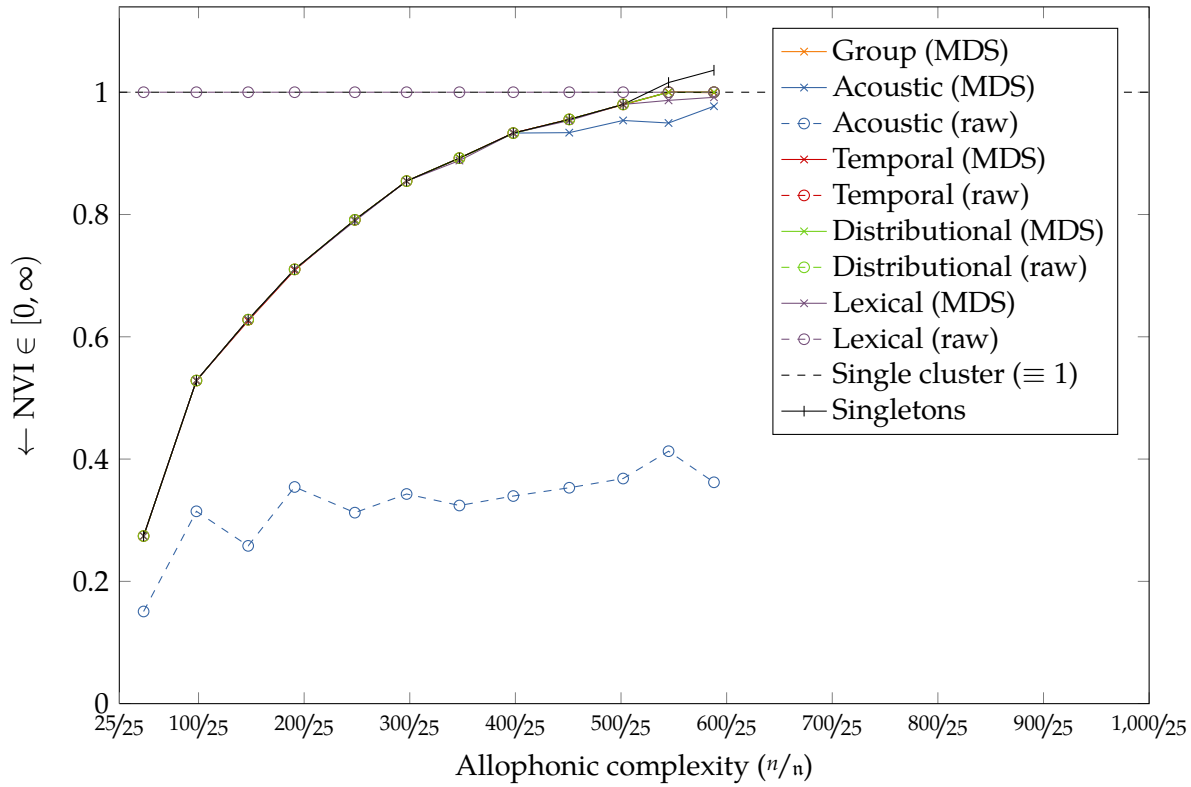


Figure 6.16 — NVI of the optimal density-based clustering solutions on the metric configurations, as a function of allophonic complexity.

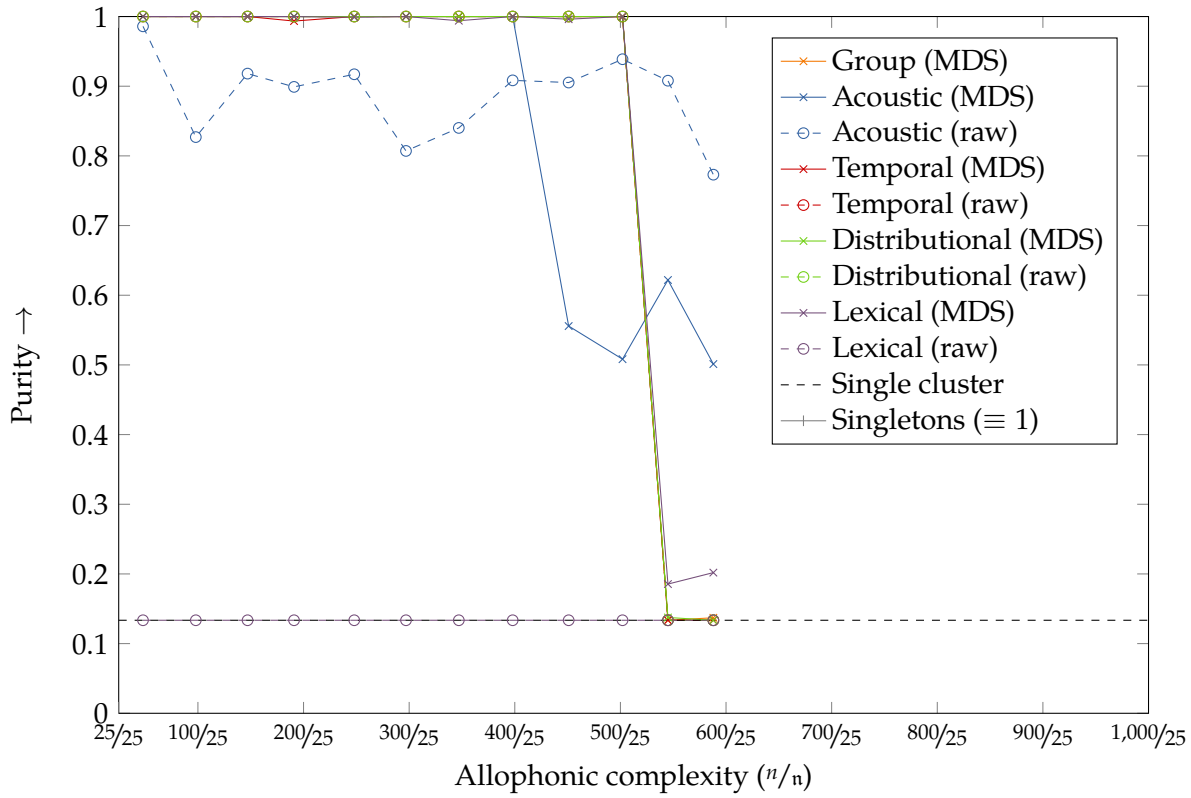


Figure 6.17 — Purity of the optimal density-based clustering solutions on the metric configurations, as a function of allophonic complexity.

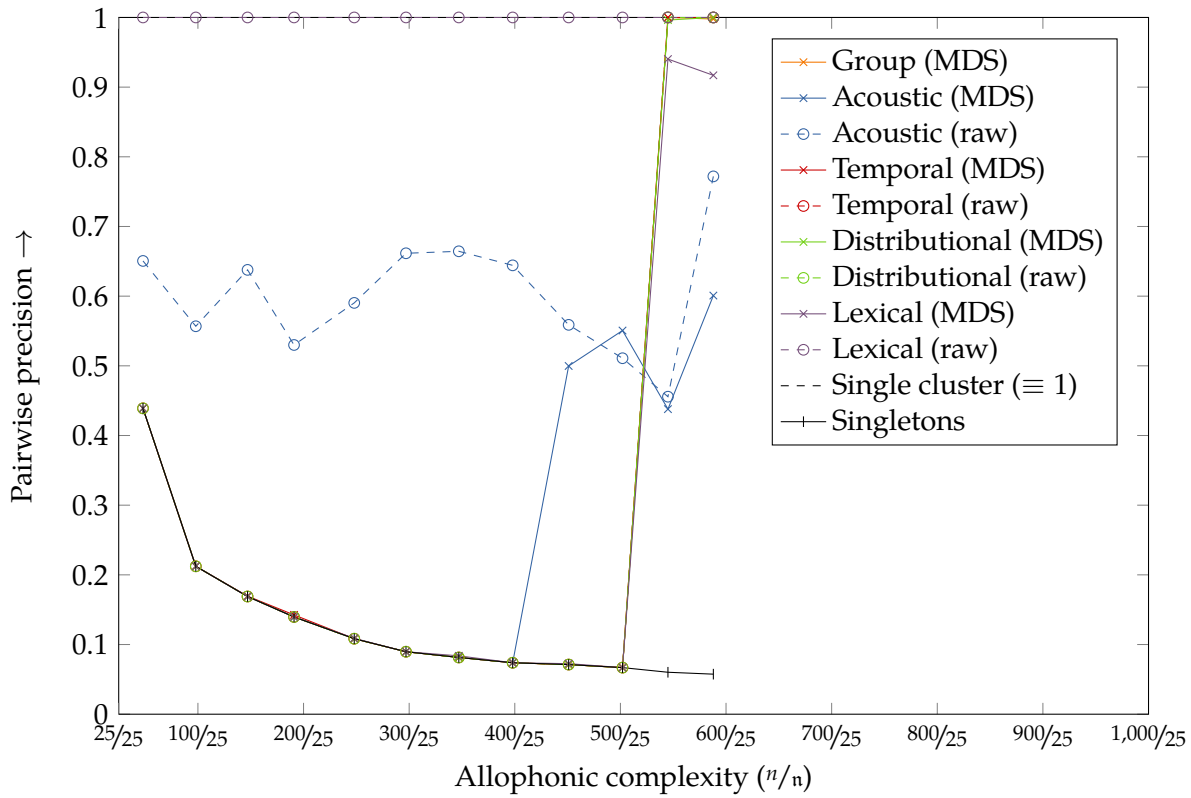


Figure 6.18 — Pairwise precision of the optimal density-based clustering solutions on the metric configurations, as a function of allophonic complexity.

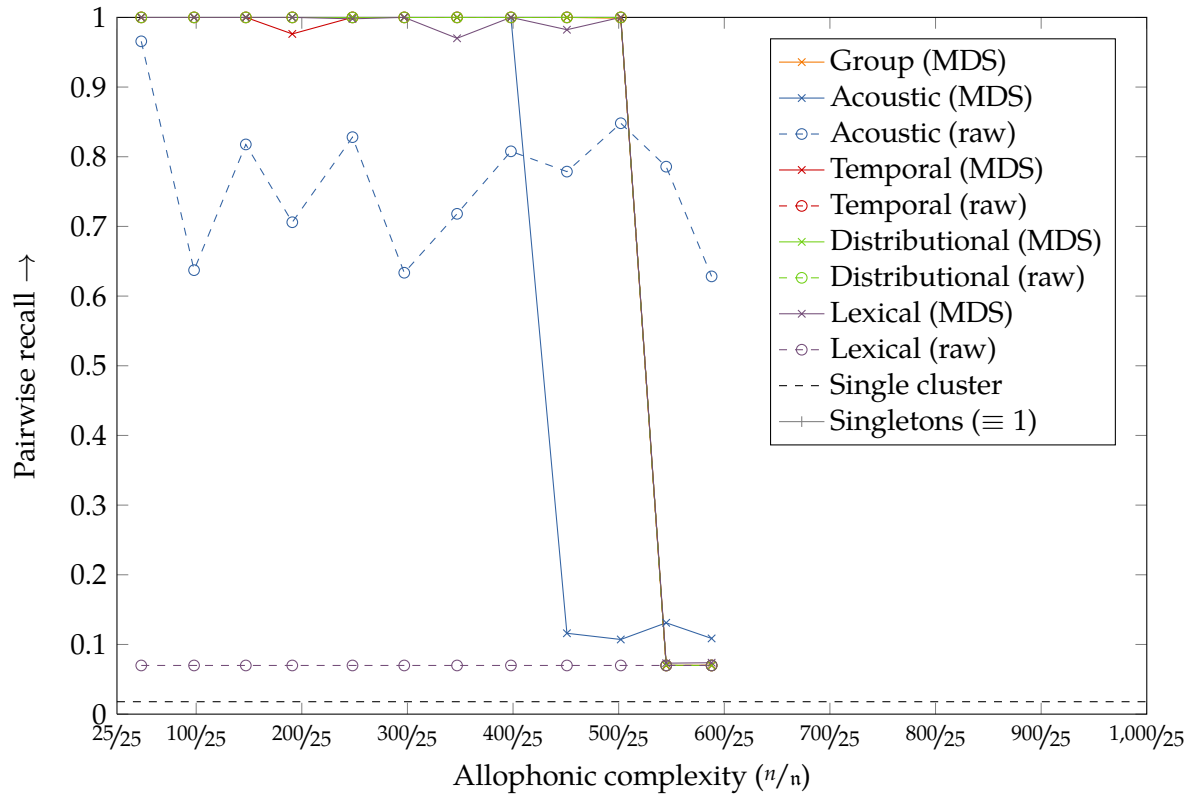


Figure 6.19 — Pairwise recall of the optimal density-based clustering solutions on the metric configurations, as a function of allophonic complexity.

As far as the acoustic indicator is concerned, comparing the NVI and purity values indicate that the density-based clustering solutions for this indicator also tend to the singletons-solutions. Indeed, whereas purity values indicate that these clustering solutions are very homogeneous—regardless of the allophonic complexity of the input—NVI values indicate that they are not as good as the reference solution. Such discrepancies between purity and NVI measurements arise from the fact that the former only evaluates homogeneity, whereas the latter evaluates both homogeneity and completeness. Acoustics-based clustering solutions are thus very homogeneous but not very complete. In other words, each cluster tends to comprise allophones of a single phoneme, but not all allophones of a given phoneme tend to belong the same cluster. Surprisingly, it is also worth noting that the Euclidean embedding of 𝔸-DTW only yields high-quality clustering solutions from $^{545}/_{25}$ allophones per phoneme onwards. This result contradicts our recurrent observation that the performance of the models decreases as the allophonic complexity of their input increases—we are unfortunately unable to account for this fact.

Altogether, the density-based clustering results confirm the prognoses of phonemehood reported in Section 6.2 and the classification results presented in Section 6.3: effective indicators of allophony are not necessarily effective indicators of phonemehood. Be that as it may, these results also confirm the relevance of MDS in the interest of embedding and combining indicators of allophony in a Euclidean space. Indeed, we showed that the 30-dimensional metric configurations may even recover some latent phonemehood-related information from the non-metric phone-to-phone dissimilarities.

6.5 Predicting phonemehood: $\mathfrak{n}$-ary clustering task

Up to this point, all classification and clustering experiments reported in this dissertation have demonstrated that effective indicators of allophony are not effective indicators of phonemehood. Moreover, we consider that enough evidence has been gathered to substantiate this claim. For these reasons, the final experiment reported in this section is *not* another attempt at recovering the phonemic inventory of Japanese from the data we introduced in Chapter 4. Here, we are interested in assessing whether—beyond phonemehood—linguistically motivated patterns emerge from a bottom-up, hierarchical clustering of the phonemic inventories.

6.5.1 Chances of phonemehood

Before turning to the actual clustering experiment, let us emphasize how daunting the task of learning a phonemic inventory is—even under the assumption of a quantized input, and when the task is formulated as a standard partitioning-clustering task.

In combinatorics, the number of distinct ways to partition a set of n labelled objects into η non-empty (unlabelled) clusters is given by the following *Stirling number of the second kind* (Graham et al., 1988; Steinley, 2007):

$$\left\{ n \atop \eta \right\} \equiv \frac{1}{\eta!} \sum_{o=0}^{\eta} (-1)^{\eta - o} \binom{\eta}{o} o^n, \tag{6.39}$$

which can be approximated (e.g. Kauffman & Rousseeuw, 1990; p. 195) by

$$\left\{ n \atop \eta \right\} \approx \frac{\eta^n}{\eta!}. \tag{6.40}$$

Assuming the optimal or desired number of clusters is known to learner (which is non-trivial), we have $\eta = \mathfrak{n} = 25$. Even for the lowest allophonic complexities $n = (48, 98, 147, 191, \dots)$, the number of possible $\mathfrak{n}$-ary clustering solutions is considerable, viz. $\left\{ n \atop \mathfrak{n} \right\} \approx (10^{40}, 10^{111}, 10^{179}, 10^{239}, \dots)$. Although dynamic programming techniques have been used for small values of n (i.e. $n \leq 30$; van Os, 2000; Hubert et al., 2001), such figures are obviously too high to enumerate all of the possible clustering partitions.

When the number of cluster is unknown, the number of distinct ways to partition a set of n labelled objects into non-empty (unlabelled) clusters is known as the n-th *Bell number*, denoted B_n. Such a number is defined as the sum of all Stirling numbers of the second kind from 0 to n (Graham et al., 1988; p. 359), i.e.

$$B_n \equiv \sum_{\eta=0}^{n} \left\{ {n \atop \eta} \right\}. \tag{6.41}$$

One can easily see from this last equation that such figures grow even faster than the corresponding Stirling numbers.

The purpose of this brief analysis is to emphasize the fact that partitioning-clustering—and, generally speaking, unsupervised learning—is a hard task. Throughout the present study, we have further assumed that only one partition is correct, viz. $\mathfrak{P}$. Consequently, for the lowest allophonic complexities $n = (48, 98, 147, 191, \dots)$, the odds of randomly discovering the phonemic partition are virtually null, viz. $\left\{ {n \atop \mathfrak{n}} \right\}^{-1} \approx (10^{-40}, 10^{-111}, 10^{-179}, 10^{-239}, \dots)$. In light of the aforestated prognosis, classification, and clustering results, we can assert that the data at hand (i.e. acoustic, temporal, distributional, and lexical indicators) are simply insufficient for the acquisition of phonemes.

$\mathfrak{n}$-means clustering algorithm It is worth noting that we did, however, perform additional partitioning-clustering experiments, using the seminal K-means clustering algorithm (algorithm by Lloyd, 1957; Forgy, 1965; MacQueen, 1967; and Hartigan & Wong, 1979; implementation by Pardo & DelCampo, 2007), and setting the desired number of clusters K to the true number of phonemes $\mathfrak{n}$. As emphasized by Mirkin (2005),

> ❝[the underlying model is] based on the somewhat simplistic claim that entities can be represented by their cluster's centroids, up to residuals.❞

Although provided with the optimal number of clusters, $\mathfrak{n}$-means clustering is indeed simplistic to discover the phonemic inventory of the target language. Compared to the previously reported density-based clustering experiments, the use of $\mathfrak{n}$-means yielded no significantly different results, which—for the sake of brevity—are not reported here.

6.5.2 Complete-linkage hierarchical clustering

In order to assess whether linguistically relevant patterns, e.g. the pervasive consonants vs. vowels dichotomy, may emerge from our data, we performed agglomerative (i.e. bottom-up) hierarchical clustering on the metric configurations.

Agglomerative clustering Hierarchical clustering techniques differ from partitioning-clustering techniques in the sense that the clustering solution the former techniques output do not consist in a partition of the input data, but in a dendrogram—a tree structure whose leaves contains the input objects, and whose branching nodes are to be interpreted as traces of the successive clustering decisions (splits or merges) taken to produce the tree.

As far as agglomerative clustering is concerned, all techniques start from the singletons solutions, i.e.

$$\Omega = \{\{p_i\} : p_i \in P\}. \tag{6.42}$$

Clusters are then iteratively merged until only one large cluster remains which contains all the observations (R Development Core Team, 2010; cf. `stats::hclust`). At each stage, the two nearest clusters in Ω are combined to form one larger cluster. In this context, various formal definitions of the word *nearest*—referred to as linkage functions—are possible. We hereby briefly review the most popular linkage functions.

The most intuitive linkage function is known as single-linkage, and is given by

$$\text{linkage}(\omega, \omega'; \mathbf{M}) \equiv \min \left\{ d_{ij}(\mathbf{M}) : p_i \in \omega, p_j \in \omega' \right\}. \tag{6.43}$$

This strategy is to be interpreted as a friends-of-friends approach whereby the distance between two clusters is computed as the distance between the two closest phones in the two clusters. An undesirable side-effect of using single-linkage—referred to as the chaining effect—is that clusters may be forced to merge due to single phones being close to one another, even if most phones in each cluster are very distant to each other (e.g. Lupşa, 2005; Manning et al., 2008).

Ward's criterion is another classic linkage function (Ward, 1963; Mirkin, 2005). Formally, Ward's criterion between two clusters is given by

$$\text{linkage}(\omega, \omega'; \mathbf{M}) \equiv \frac{|\omega|\,|\omega'|}{|\omega| + |\omega'|}\, d\left(\bar{\omega}, \bar{\omega}'\right) \quad \text{where} \quad \bar{\omega} \equiv \frac{1}{|\omega|} \sum_{p_i \in \omega} \mathbf{m}_i. \tag{6.44}$$

It is worth noting, however, that using this linkage function is unsuitable in the case at hand, as it relies on the centroids $\bar{\omega}$ and $\bar{\omega}'$. We have indeed shown in Sections 6.2 and 6.3 that our data violate the compactness hypothesis, as the true phonemic clusters are neither compact nor well-separated. This approach would thus suffer the same limitations as K-means (Mirkin, 2005).

Finally, the linkage function we use in this experiment is known as the complete-linkage. In this approach to agglomerative clustering, the distance between two clusters is computed as the distance between the two farthest phones in the two clusters, i.e.

$$\text{linkage}(\omega, \omega'; \mathbf{M}) \equiv \max\left\{d_{ij}(\mathbf{M}) : p_i \in \omega, p_j \in \omega'\right\}. \tag{6.45}$$

Complete-linkage tends to produce tightly bound clusters and, to our knowledge, has no known undesirable side-effect (Lupşa, 2005).

Phonemic cut Hierarchical clustering techniques yield dendrograms with as many leaves as there are phones in the input allophonic inventory. For the sake of simplicity, we therefore cut each dendrogram so as to obtain a canopy-clustered dendrogram with $\mathfrak{n} = 25$ leaves only.

6.5.3 Evaluation

For the purpose of the present analysis, no quantitative criteria appeared to be appropriate. We therefore limited the evaluation of the hierarchical clustering solutions to a visual examination of the output dendrograms. Because such dendrograms are space-consuming data representations, we further limited our examination to the group configurations of the combined model.

6.5.4 Results

The resulting dendrograms for the group configurations are presented in Figures 6.20, 6.21, 6.22, 6.23, 6.24, and 6.25 at allophonic complexities of $^{48}/_{25}$, $^{98}/_{25}$, $^{191}/_{25}$, $^{347}/_{25}$, $^{502}/_{25}$, and $^{588}/_{25}$ allophones per phoneme, respectively. In each dendrogram, branches are ordered so that, for each internal node, the tighter cluster is on the left. Subscript figures indicate, for each phoneme, the percentage of its (weighted) occurrences falling in that cluster; ε denotes quantities less than a percent. Components may not sum to totals because of rounding.

Unfortunately, no definite pattern emerges in these dendrograms—regardless of the allophonic complexity of the input. On one hand, many leaf-clusters are individually complete and homogeneous. On the other hand, all dendrograms contain one or more catch-all clusters wherein most—if not all—occurrences of varied phonemes are aggregated. Moreover, no natural class is further revealed by the arborescent structure of the dendrograms. The ubiquitous dichotomy between consonants and vowels, for example, is never accounted for by the higher-level branching nodes of the dendrograms. All in all, these additional negative results confirm our recurrent conclusion throughout this chapter: effective indicators of allophony are not effective indicators of phonemehood.

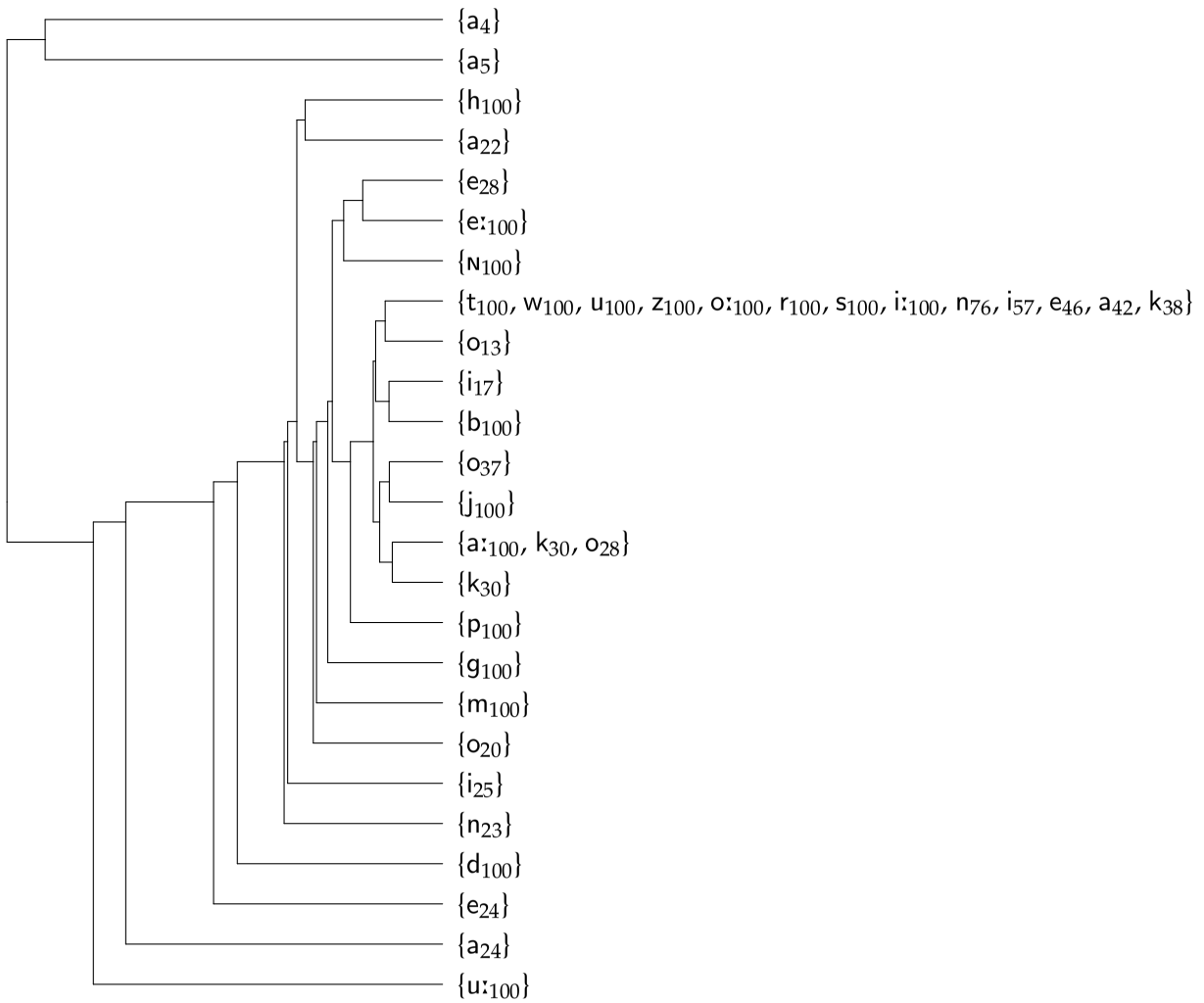


Figure 6.20 — Dendrogram for the complete-linkage hierarchical clustering of the group metric configuration at $n/\mathfrak{n} = 48/25$ allophones per phoneme, cut at $\mathfrak{n}$ clusters. Branches are ordered so that, for each internal node, the tighter cluster is on the left. Subscript figures indicate, for each phoneme, the percentage of its (weighted) occurrences falling in that cluster. Components may not sum to totals because of rounding.

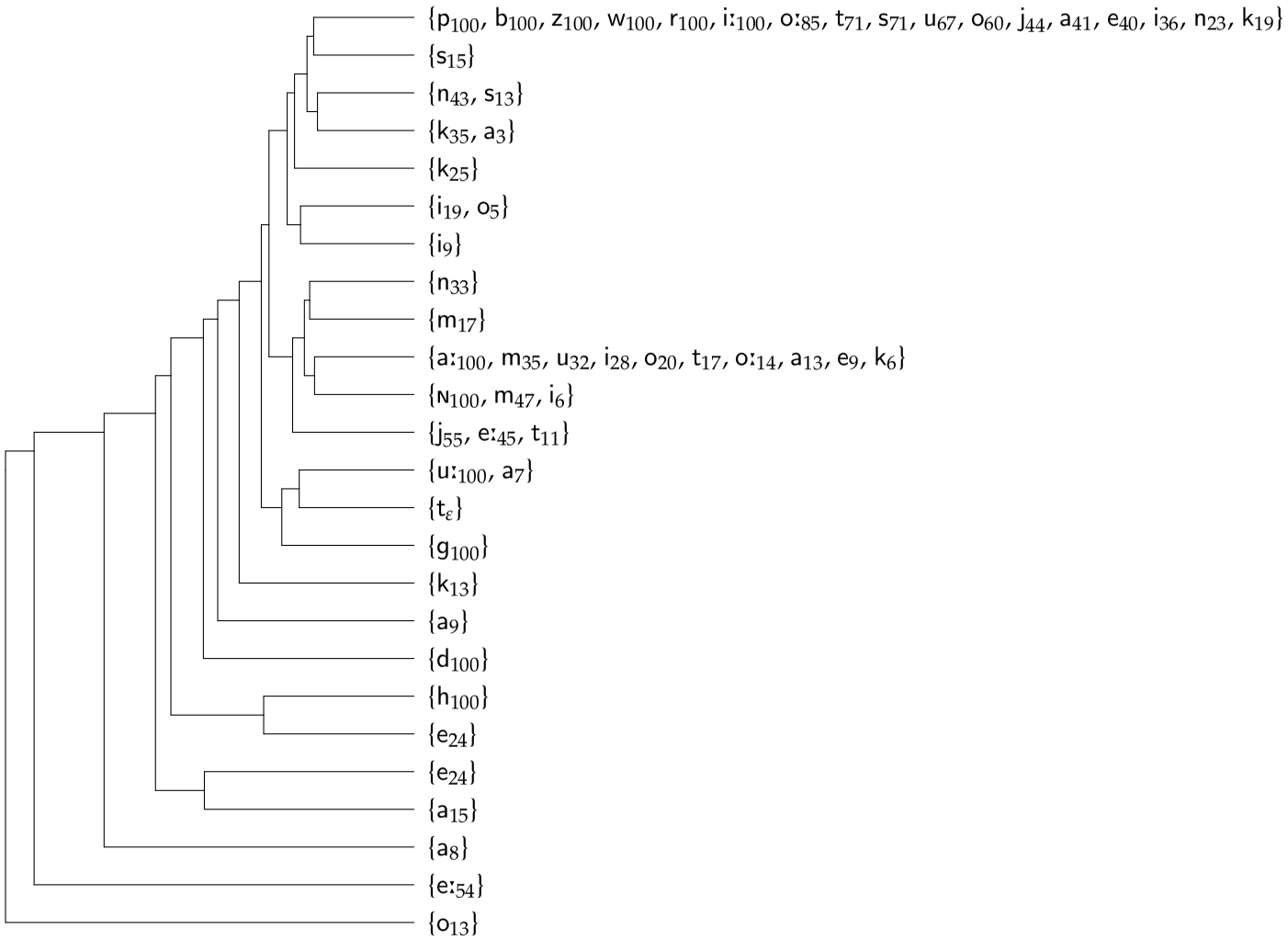


Figure 6.21 — Dendrogram for the complete-linkage hierarchical clustering of the group metric configuration at $^{n}/_{\mathfrak{n}} = {}^{98}/_{25}$ allophones per phoneme, cut at $\mathfrak{n}$ clusters. Branches are ordered so that, for each internal node, the tighter cluster is on the left. Subscript figures indicate, for each phoneme, the percentage of its (weighted) occurrences falling in that cluster. Components may not sum to totals because of rounding.

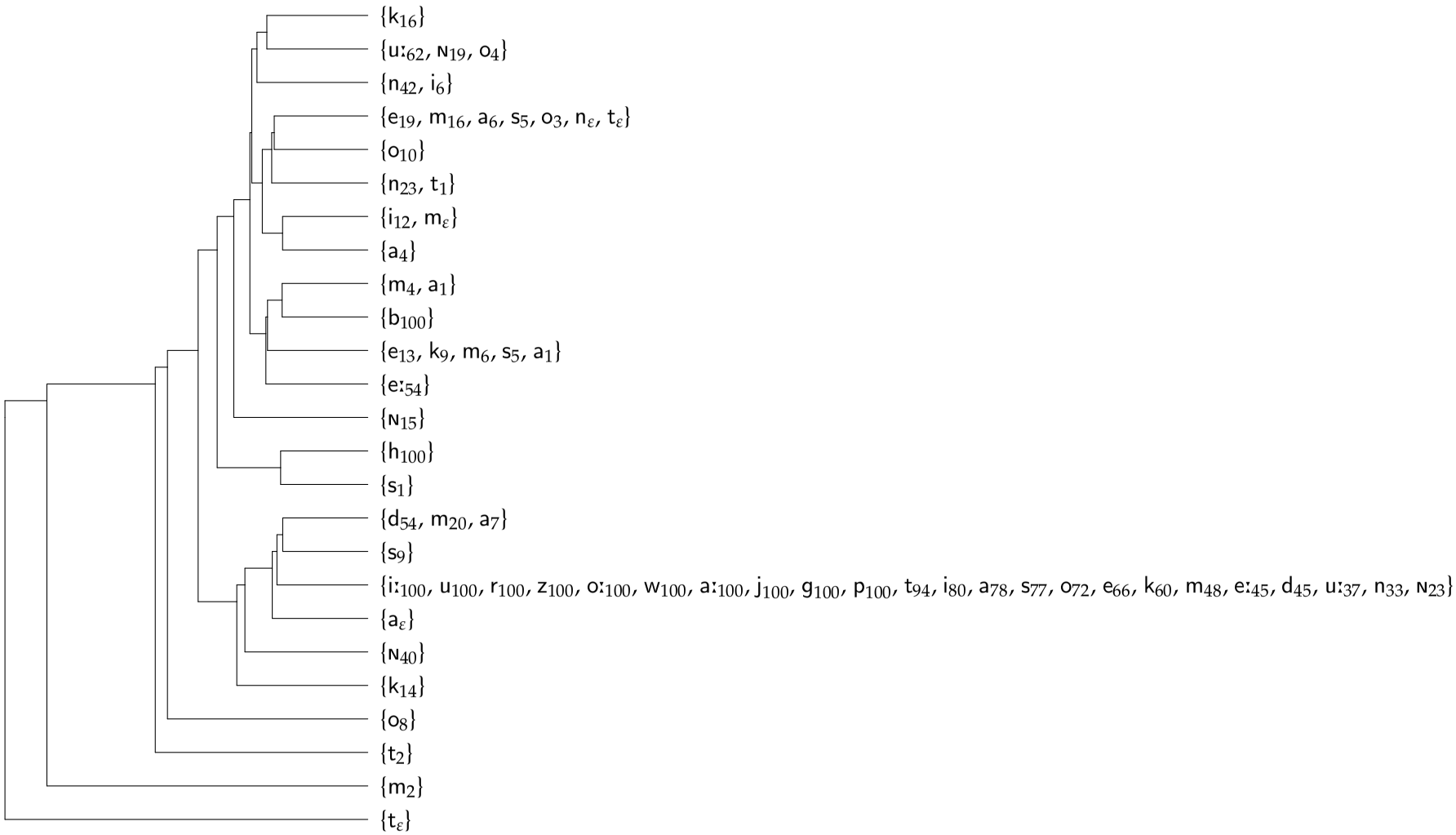


Figure 6.22 — Dendrogram for the complete-linkage hierarchical clustering of the group metric configuration at $^{n}/_{\mathfrak{n}} = {}^{191}/_{25}$ allophones per phoneme, cut at $\mathfrak{n}$ clusters. Branches are ordered so that, for each internal node, the tighter cluster is on the left. Subscript figures indicate, for each phoneme, the percentage of its (weighted) occurrences falling in that cluster. Components may not sum to totals because of rounding.

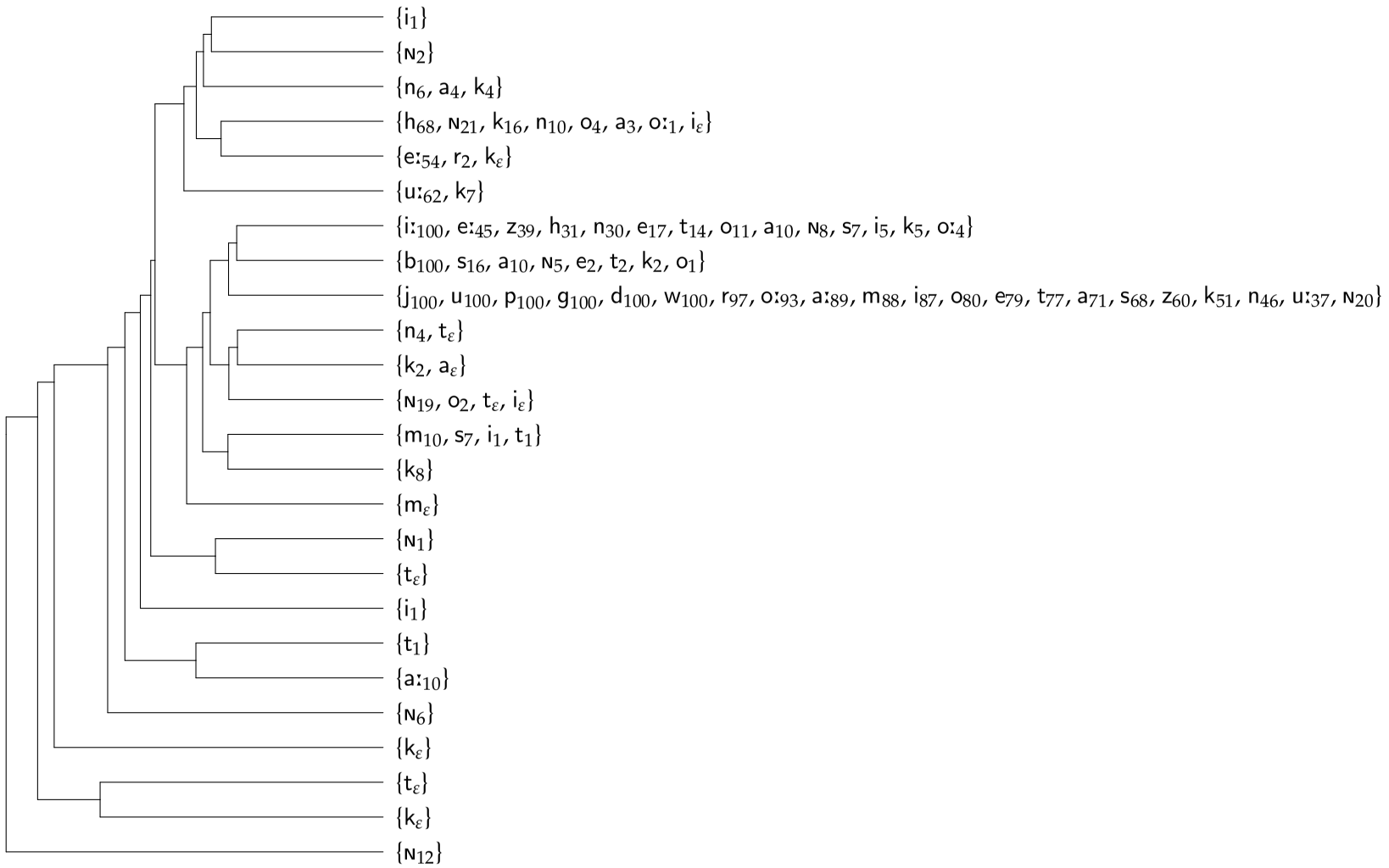


Figure 6.23 — Dendrogram for the complete-linkage hierarchical clustering of the group metric configuration at $^{n}/_{\mathfrak{n}} = {}^{347}/_{25}$ allophones per phoneme, cut at $\mathfrak{n}$ clusters. Branches are ordered so that, for each internal node, the tighter cluster is on the left. Subscript figures indicate, for each phoneme, the percentage of its (weighted) occurrences falling in that cluster. Components may not sum to totals because of rounding.

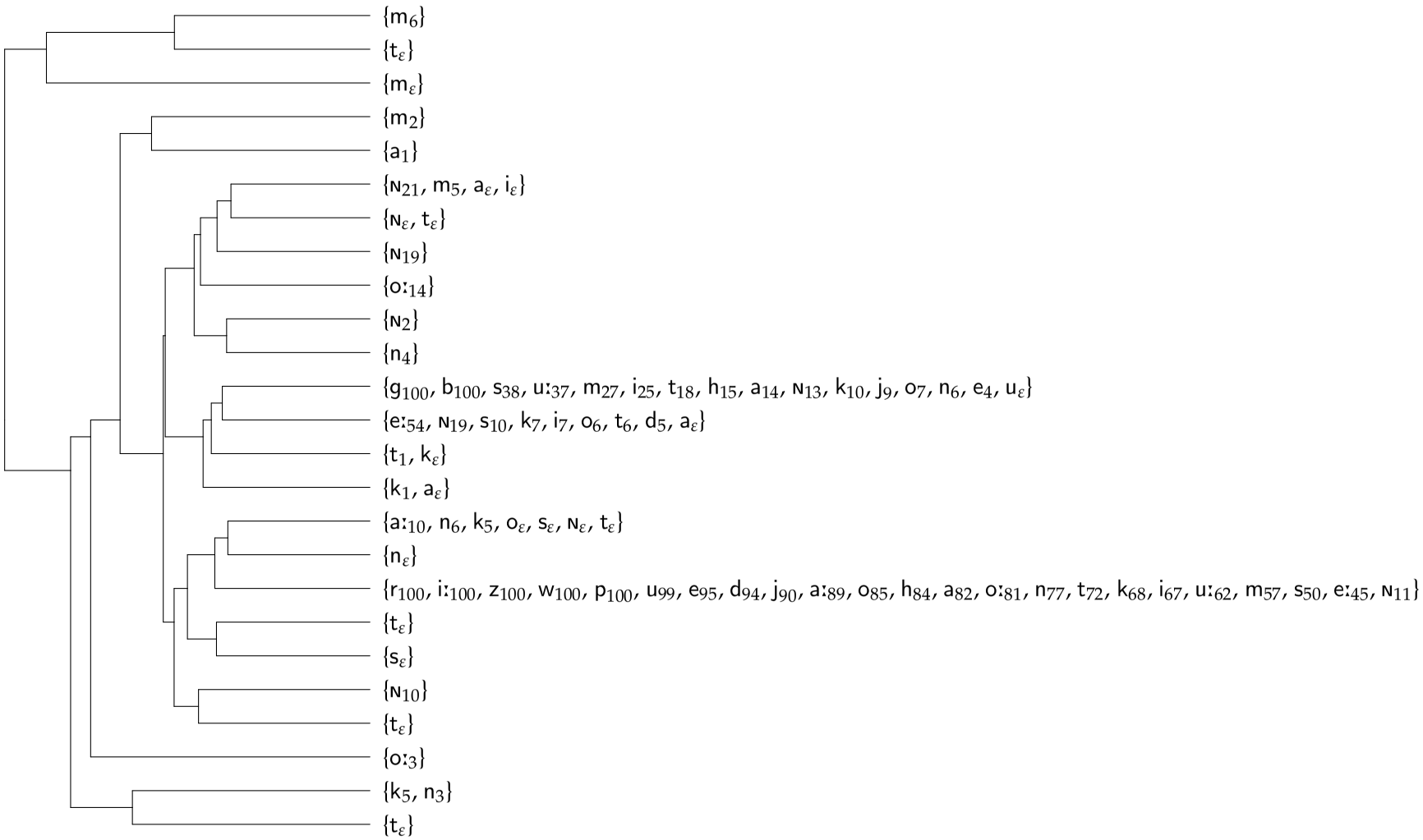


Figure 6.24 — Dendrogram for the complete-linkage hierarchical clustering of the group metric configuration at $^{n}/_{\mathfrak{n}} = {}^{502}/_{25}$ allophones per phoneme, cut at $\mathfrak{n}$ clusters. Branches are ordered so that, for each internal node, the tighter cluster is on the left. Subscript figures indicate, for each phoneme, the percentage of its (weighted) occurrences falling in that cluster. Components may not sum to totals because of rounding.

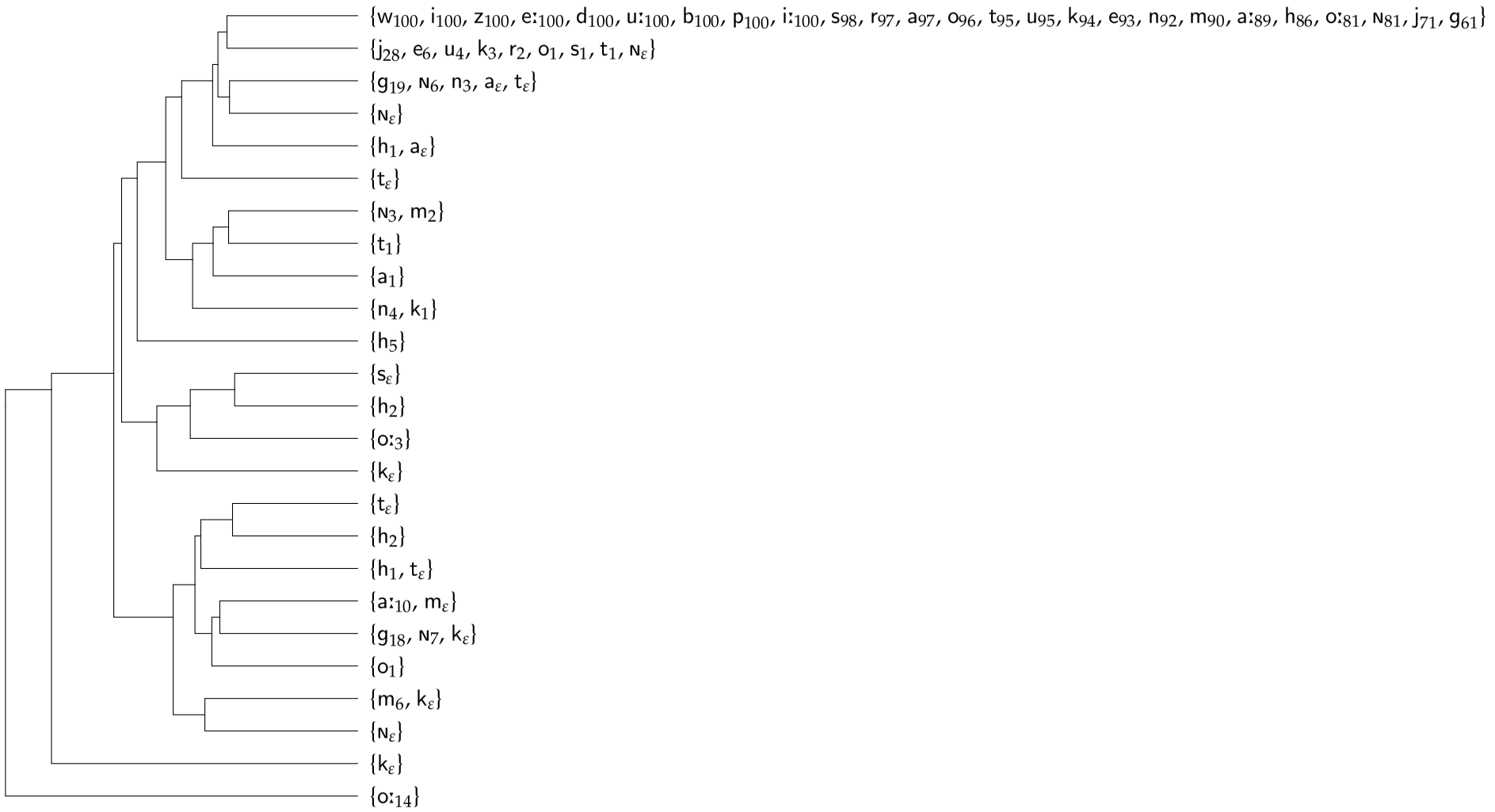


Figure 6.25 — Dendrogram for the complete-linkage hierarchical clustering of the group metric configuration at $^{n}/_{\mathfrak{n}} = {}^{588}/_{25}$ allophones per phoneme, cut at $\mathfrak{n}$ clusters. Branches are ordered so that, for each internal node, the tighter cluster is on the left. Subscript figures indicate, for each phoneme, the percentage of its (weighted) occurrences falling in that cluster; ε denotes quantities less than a percent. Components may not sum to totals because of rounding.

6.6 Overall assessment

In this chapter, we presented the major contribution of the present study, i.e. a complete reformulation of Peperkamp et al.'s (2006) framework based on three-way MDS. Given three-way input dissimilarities, MDS models yield metric configurations in Euclidean spaces of arbitrary dimensionality, so that the point-to-point distances in that space account for the input phone-to-phone dissimilarities. We further showed in Section 6.4 that such Euclidean embeddings of our indicators of allophony do not smear the information contained in the indicators.

The main conclusions of this chapter, however, that indicators of allophony are not indicators of phonemehood—as uniformly showed by the various prognosis, classification, and clustering results presented in Sections 6.2, 6.3, 6.4, and 6.5. Although such results are, at a first glance, disappointing, we argued in Section 6.5.1 that the odds of randomly discovering the reference phonemic partition $\mathfrak{P}$ are virtually null.

Further research is thus needed in order to develop indicators that are not only strongly correlated with allophony, but also with phonemehood, i.e. with the identity of phonemes.

CHAPTER 7
CONCLUSION

The pervasiveness of allophonic processes in natural languages is undeniable. In light of prior scholarship and of the experiments presented throughout this dissertation, the question of how infants may acquire the phonemic inventory of their native language remains an open question. The general conclusion of our work is indeed that no acoustic, temporal, distributional, or lexical indicator of allophony (or any combination thereof) appears to contain enough information for the discovery of well-separated, phoneme-like sound categories.

In this final chapter, we review the scientific question motivating the present study, as well as our contribution to knowledge. We then go on to discuss recommendations for future research, notably in view of the various linguistic and computational limitations of our work.

7.1 Indicators of allophony and phonemehood

This dissertation describes a modeling project aimed at assessing the relevance and effectiveness of various linguistic cues, in the interest of reducing the considerable, inherent variability in speech to a small number of sounds categories, i.e. the phonemes of the target language. Throughout the present study, we assumed the position of a naive learner—more precisely, that of an infant acquiring its native language. Indeed, having observed that the realizations of a given phoneme are constrained by the surrounding phonemes (or the realizations thereof, cf. Figure 2.2), one can but be puzzled by the apparent ease with which infants bootstrap into language. How, for instance, do Japanese-learning infants discover that the third sounds in [miso] ∼ [moʃi] (*miso* ∼ *moshi*) are both to be processed as realizations of the same abstract sound category? What kind of cues may English-learning infants rely on to learn that [sɪŋkɪŋ] ∼ [θɪŋkɪŋ] (*sinking* ∼ *thinking*) can not refer to the same action?

The work presented in this dissertation builds upon the line of computational studies initiated by Peperkamp et al. (2006). The common hypothesis underlying this body of work is that infants are able to keep track of—and rely on—phone-to-phone dissimilarity judgments in order to eventually cluster similar phones into phonemic categories. In Chapter 4, we showed that all previously proposed phone-to-phone dissimilarities can actually be reformulated as various instances of a single, cohesive concept, viz. indicators of allophony. In that chapter, we then benchmarked the (relative) efficiency of indicators of allophony that rely on different types of linguistic cues, i.e. measures of near-complementary distributions (Peperkamp et al., 2006; Le Calvez et al., 2007), acoustic similarity (Dautriche, 2009), lexical dissimilarity (Martin et al., 2009; Boruta, 2011b), and temporal similarity—a novel class of indicators introduced in this study. To this aim, we amended Peperkamp et al.'s (2006) artificial learner in order to obtain empirical upper bounds on the learnability of the allophony relation, i.e. the binary relation that brings together phones that are realizations of the same phoneme. Throughout the extensive evaluation carried out in Chapter 4, we showed that indicators' effectiveness at distinguishing between

Table 7.1 — Classification and clustering experiments reported throughout this dissertation.

Supervision	Task	Input	Algorithm	Section
Supervised	Binary	$n \times n \times \kappa$ proximity matrix	Binomial regression	4.5
Supervised	$(\mathfrak{n}+1)$-ary	$n \times n \times \kappa$ proximity matrix	Multinomial regression	5.2
Supervised	$\mathfrak{n}$-ary	$n \times q$ metric configuration	Multinomial regression	6.3
Unsupervised	?-ary	$n \times q$ metric configuration	Density-based clustering	6.4
Unsupervised	$\mathfrak{n}$-ary	$n \times q$ metric configuration	Hierarchical clustering	6.5

allophonic and non-allophonic pairs of phones is mostly due to the correct rejection of the latter. As illustrated in the confusion plots we introduced in Section 4.4, this (undesirable) effect is due to the fact that—in any given allophonic inventory—allophonic pairs are both statistically outnumbered and distributionally overwhelmed by non-allophonic pairs.

In Chapter 5, we further addressed the impediments of Peperkamp et al.'s (2006) framework for the discovery of phonological knowledge, and argued that the major limitations arise from its pairwise formulation. Because the elementary object to be considered by Peperkamp et al.'s learner is a pair of phones, the input representation of the data is smeared by the ubiquity of non-allophonic pairs, which—as we have argued throughout this dissertation—are simply the spurious byproduct of a combinatorial enumeration. Indeed, the polyadic phonemehood relation can be fully accounted for by the allophonic pairs alone and, hence, non-allophonic pairs are—to say the least—redundant. Furthermore, distributional and lexical indicators of allophony require that the input be quantized, as argued in Section 3.2. Consequently, if a given (low-variability) phoneme is assigned a single representation in the allophonic inventory, this phoneme can not appear in any allophonic pair and, hence, is merely cancelled out in such a pairwise framework. Finally, we showed that all studies building upon Peperkamp et al.'s (2006) original experiments made contradictory modeling assumptions, viz. that pairs of phones are statistically independent and, indirectly, that allophony is a transitive relation. As we argued in Section 5.1.1, these working hypotheses are contradictory, as the former negates the underlying structure in the data that the latter demands. To us, all aforestated limitations call for a reformulation of Peperkamp et al.'s framework wherein the phonemic inventory of the target language can be learned directly, i.e. without the intermediate learning of the allophony relation. Unpairing phone pairs requires a non-trivial reformulation of the input data and, hence, impedes the comparability between previously published studies and the experiments reported in this dissertation. For this reason, we carried out transitional experiments whereby only one feature of the whole model (task, input, or supervision) differed from one experiment to the next—as illustrated in Table 7.1. However, the resuts of our preliminary attempts at learning phoneme-like sound categories were disappointing (cf. Section 5.2), as only the most frequent phonemes were recovered from the phone-to-phone dissimilarities contained in our indicators of allophony.

In Chapter 6, we presented the major contribution of the present study, i.e. a complete reformulation of Peperkamp et al.'s (2006) framework based on a statistical technique known as multidimensional scaling (Torgerson, 1952; Groenen & van de Velden, 2004; Borg & Groenen, 2005). Let us emphasize the most attractive features of this technique:

— according to this model, the mind generates an impression of dissimilarity by adding up the perceived differences of two phones over their properties;
— the embedding of the input phone-to-phone dissimilarities occurs in a sound and standard Euclidean space of arbitrary dimensionality;
— the embedding yields workable coordinates for every phone in the allophonic inventory;
— three-way extensions are readily available, allowing us to address simultaneously the issues of embedding and combining of indicators of allophony.

Though appealing, it is worth mentioning that this technique is computationally prohibitive as the time complexity of the optimization procedure is quadratic in the number of phones in the

allophonic inventory—to say nothing of the number of iterations, dimensions, and combined indicators. Be that as it may, we consider multidimensional scaling an effective technique, chiefly on the grounds that it allows for the use of individual phones as the elementary objects to be considered by the learner. Relying on such Euclidean embeddings of indicators of allophony (and combinations thereof), we further argued in Chapter 6 that the task of learning the phonemic inventory of the target language can in fact be recast as a standard partitioning-clustering task, whereby the learner has to partition a set of phones into collectively exhaustive and mutually exclusive phonemes. We then carried out various classification and clustering experiments whose results consistently indicate that effective indicators of allophony are not necessarily effective indicators of phonemehood. Indeed, none of the indicators of allophony we considered in the present study appears to provide compact or well-separated phoneme-like clusters. In other words, our input data violate the compactness hypothesis that underlies most clustering problems (Duin, 1999; Pękalska et al., 2003) and, as we argued in Section 2.1, early language acquisition. Although attempting to learn the phonemehood relation from such data thus appears to be in vain, it is worth highlighting our explanation in Section 6.4 that the mediocre performance we observed throughout Chapter 6 is not the consequence of our embedding-based reformulation of the problem: using the original phone-to-phone dissimilarities does not yield models with a better predictive power in the interest of discovering phonemes.

Altogether, the computational results we discussed throughout this dissertation suggest that allophony and phonemehood can only be discovered from acoustic, temporal, distributional, or lexical indicators at low allophonic complexities, i.e. when—on average—phonemes do not have many allophones in the quantized representation of the input. Two competing explanations may account for this fact. Were infants to tackle a considerable number of allophones, the work reported in this dissertation would fail to propose a plausible account of early phonological acquisition. On the contrary, were infants to tackle a number of allophones commensurable with the allophonic complexities at which our models were evaluated, then the point at which the performance of our models suffer a significant drop would be interpreted as an assessment of the number of allophonic processes at stake in infant-directed speech.

7.2 Future research

A number of issues and limitations have been brought to light during this work. First, because of the limited availability of databases of infant-directed speech with aligned transcriptions, we have used a corpus of adult-directed speech, viz. the *Corpus of Spontaneous Japanese* (Maekawa et al., 2000; Maekawa, 2003). However, as we argued in Section 3.1, using infants' linguistic input to model infants' linguistic behavior would be a more plausible alternative. Furthermore, we consider that computational models of early language acquisition should be evaluated using data from typologically different languages in order to assess their sensitivity to linguistic diversity (Gambell & Yang, 2004; Boruta et al., 2011). We therefore leave as a recommendation for future research the task of reproducing the work presented throughout this dissertation using other corpora in order to assess whether our conclusions can be applied generally to language acquisition, or if they are mere numerical accidents due to the serendipitous, conjoint use of this particular corpus of Japanese and, for example, density-based clustering.

Moreover, we have shown in the present study that acoustic, temporal, distributional, or lexical indicators of allophony yield dissimilarity judgments whereby non-allophonic pairs of phones tend to be more dissimilar that allophonic pairs—hence confirming the relevance of Peperkamp et al.'s (2006) proposal. We have also shown that such indicators are, however, perfectible, as they do not yield an acceptable separation of the phonemes of the target language. Further research is thus needed to fully transform indicators of allophony into indicators of phonemehood. Regarding the particular framework upon which our work builds, another advancement would consist in examining the relative influence of the constraints one can put on the three-way multidimensional scaling models. In the absence of a priori linguistic knowledge—

as well as prior scholarship on the use of such models for language-related problems—we indeed favored the least constrained three-way model. Though non-trivial, we acknowledge that this choice was arbitrary and, thus, that its consequences need to be further investigated.

Finally, let us revisit a fundamental assumption of the work presented in this dissertation: the development of speech perception is an innately guided learning process (Jusczyk & Bertoncini, 1988; Gildea & Jurafsky, 1996; Plunkett, 1997; Bloom, 2010). Our results confirm this assumption in the sense that density-based clustering, i.e. the least constrained of all algorithms we used, yielded the least accurate predictions. The crucial point here is that the discovery of linguistic knowledge is facilitated—if not simply made possible—by a priori linguistic constraints. Interestingly enough, Johnson et al. (2007) recently proposed a class of probabilistic models of language, referred to as adaptor grammars, that can be constrained so as to account for some given linguistic structure. This approach has been used to develop and validate various computational models of the acquisition of word segmentation (Johnson, 2008a,b) but, to our knowledge, no experiment focusing on the acquisition of phonological categories has yet been reported. We therefore leave as a recommendation for future research the task of assessing whether providing an artificial learner with computational constraints or linguistic structure facilitates the acquisition of phonemes, as, in the words of Bloom (2010):

> ❝One lesson from the study of artificial intelligence [...] is that an empty head learns nothing: a system that is capable of rapidly absorbing information needs to have some prewired understanding of what to pay attention to and what generalizations to make. Babies might start off smart, then, because it enables them to get smarter.❞

❝Shadow didn't wait to see. He walked away and he kept on walking.❞
— Neil Gaiman, in *American Gods*

REFERENCES

Nota bene Numbers in square brackets are backreferences to the main text.

Abney, S. (2011). Data-intensive experimental linguistics. *Linguistic Issues in Language Technology 6*(2). [pp. 20 and 22]

Agresti, A. (2007). *An Introduction to Categorical Data Analysis* (second ed.). Hoboken, New Jersey: John Wiley & Sons, Inc. [pp. 23, 83, 84, 85, 88, 102, 103, and 127]

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control 19*(6), 716–723. [p. 86]

Aliguliyev, R. M. (2009). Performance evaluation of density-based clustering methods. *Information Sciences 179*, 3583–3602. [pp. 100 and 125]

Alishahi, A. (2011). *Computational Modeling of Human Language Acquisition*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool. [pp. 21, 23, and 82]

AMLaP. Architectures and mechanisms for language processing. http://www.amlap.org/. [p. 22]

Ankerst, M., M. M. Breunig, H.-P. Kriegel, and J. Sander (1999). OPTICS: ordering points to identify the clustering structure. In *Proceedings of the ACM SIGMOD International Conference on Management of data*, pp. 49–60. ACM Press. [p. 135]

Aslin, R. N., J. R. Saffran, and E. L. Newport (1998). Computation of conditional probability statistics by 8-month old infants. *Psychological Science 9*, 321–324. [pp. 19 and 24]

Atkinson, Q. D. (2011). Phonemic diversity supports a serial founder effect model of language expansion from Africa. *Science 332*(6027), 346–349. [pp. 14 and 157]

Austin, W. M. (1957). Criteria for phonetic similarity. *Language 33*(4), 538–544. [pp. 15 and 54]

Bane, M. (2011). `simplegoodturing`: a Python module implementing the "Simple Good–Turing" method of frequency estimation described by Gale & Sampson (1995), version 0.3. Retrieved October 21, 2011. https://github.com/maxbane/simplegoodturing. [p. 60]

Barzily, Z., Z. Volkovich, B. Akteke-Öztürk, and G.-W. Weber (2009). On a minimal spanning tree approach in the cluster validation problem. *Informatica 20*(2), 187–202. [p. 125]

Berdicevskis, A. and A. C. Piperski (2011). Doubts about a serial founder-effect model of language expansion. *Science*. Response to Atkinson's (2011) article. Published online 8 December 2011. http://www.sciencemag.org/content/332/6027/346.abstract/reply. [p. 22]

Bergstra, J. and Y. Bengio (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research 13*, 281–305. [pp. 134 and 135]

Bezdek, J. C. and N. R. Pal (1998). Some new indices of cluster validity. *IEEE Transactions on Systems, Man and Cybernetics 28*(3), 301–315. [pp. 100 and 125]

Bhattacharyya, A. (1943). On a measure of divergence between two statistical populations defined by their probability distributions. *Bulletin of the Calcutta Mathematical Society 35*, 99–109. [p. 59]

Bloom, P. (2010). The moral life of babies. *The New York Times Magazine*, MM44. May 9, 2010. [pp. 19, 24, and 154]

Boersma, P. (2010). Modelling phonological category learning. Accessed March 15, 2012. http://www.fon.hum.uva.nl/paul/. [pp. 21 and 24]

Boersma, P. (2011). A programme for bidirectional phonology and phonetics and their acquisition and evolution. In A. Benz and J. Mattausch (eds.), *Bidirectional Optimality Theory*, pp. 33–72. Amsterdam: John Benjamins. [p. 21]

Boersma, P., P. Escudero, and R. Hayes (2003). Learning abstract phonological from auditory phonetic categories: an integrated model for the acquisition of language-specific sound categories. In *Proceedings of the Fifteenth International Congress of Phonetic Sciences*, Barcelona, pp. 1013–1016. [p. 21]

Boersma, P. and C. Levelt (2000). Gradual constraint-ranking learning algorithm predicts acquisition order. In *Proceedings of Thirtieth Child Language Research Forum*, Stanford, California, pp. 229–237. [p. 21]

Boersma, P. and C. Levelt (2003). Optimality theory and phonological acquisition. Volume 3, pp. 1–50. [p. 21]

Borg, I. and P. J. F. Groenen (2005). *Modern Multidimensional Scaling: Theory and Applications* (second ed.). Springer Series in Statistics. [pp. 116, 117, 118, 119, 120, 121, and 152]

Boruta, L. (2009). Acquisition précoce des règles allophoniques: approche computationnelle. Master's thesis, Université Paris Diderot. [pp. 21, 22, 26, 33, 34, 51, 61, 62, 66, 75, 82, 94, 95, 98, and 106]

Boruta, L. (2011a). A note on the generation of allophonic rules. Technical Report #0401, INRIA. [pp. 26, 33, 34, and 35]

Boruta, L. (2011b). Combining indicators of allophony. In *Proceedings of the ACL 2011 Student Session*, Portland, Oregon, pp. 88–93. Association for Computational Linguistics. [pp. 16, 21, 22, 26, 33, 34, 51, 53, 59, 61, 62, 63, 64, 66, 69, 74, 75, 76, 82, 87, 88, 95, 98, 106, and 151]

Boruta, L., S. Peperkamp, B. Crabbé, and E. Dupoux (2011). Testing the robustness of online word segmentation: effects of linguistic diversity and phonetic variation. In *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics*, Portland, Oregon, pp. 1–9. Association for Computational Linguistics. [pp. 20, 21, 26, 33, 34, 62, and 153]

Brannen, K. (2002). The role of perception in differential substitution. *The Canadian Journal of Linguistics 47*(1/2), 1–46. [p. 14]

Brent, M. R. (1999). An efficient, probabilistically sound algorithm for segmentation and word discovery. *Machine Learning 34*(1–3), 71–105. [pp. 20, 21, 61, 82, and 166]

Brent, M. R. and T. A. Cartwright (1996). Distributional regularity and phonotactic constraints are useful for segmentation. *Cognition 61*, 93–125. [pp. 20 and 61]

Bronstein, M. M., A. M. Bronstein, R. Kimmel, and I. Yavneh (2005). A multigrid approach for multi-dimensional scaling. In *Proceedings of the Twelfth Copper Mountain Conference on Multigrid Methods*. [p. 117]

Carroll, J. D. and J.-J. Chang (1970). Analysis of individual differences in multidimensional scaling via an *n*-way generalization of "Eckart-Young" decomposition. *Psychometrika 35*(2), 283–319. [p. 119]

Chen, X., P. Ender, M. Mitchell, and C. Wells (2012). Logistic regression with Stata. UCLA: Academic Technology Services, Statistical Consulting Group. Accessed April 15, 2012. http://www.ats.ucla.edu/stat/stata/webbooks/logistic/. [pp. 83, 84, and 85]

Chomsky, N. and M. Halle (1968). *The Sound Pattern of English* (1991 ed.). The MIT Press. [pp. 14, 16, 21, 35, and 45]

Chou, C.-H., M.-C. Su, and E. Lai (2004). A new cluster validity measure and its application to

image compression. *Pattern Analysis and Applications 7*, 205–220. [p. 125]

Christiansen, M. H., J. Allen, and M. Seidenberg (1998). Learning to segment speech using multiple cues. *Language and Cognitive Processes 13*(2), 221–268. [pp. 20 and 61]

Clapper, B. M. (2009). `munkres`: a Python module implementing the "Hungarian method" described by Munkres (1957). Version 1.0.5.3. http://software.clapper.org/munkres/. [p. 56]

Clark, E. V. (2004). How language acquisition builds on cognitive development. *Trends in Cognitive Sciences 8*(10), 472–478. [p. 18]

CMCL. Cognitive modeling and computational linguistics. http://www.psy.cmu.edu/~cmcl/. [p. 22]

Coleman, J. (1998). *Phonological representations: their names, forms, and powers*. Cambridge Studies in Linguistics 85. Cambridge: Cambridge University Press. [pp. 13, 14, and 24]

Coolen, A. C. C., R. Kühn, and P. Sollich (2005). *Theory of Neural Information Processing Systems*. Oxford: Oxford University Press. [p. 63]

Cormen, T. H., C. E. Leiserson, R. L. Rivest, and C. Stein (2001). *Introduction to Algorithms* (second ed.). The MIT Press. [pp. 56 and 98]

Cover, T. M. and J. A. Thomas (2006). *Elements of Information Theory* (second ed.). Wiley–Interscience. [p. 59]

Cristia, A. (2011). Ongoing meta-analysis on the acoustics of infant-directed speech. Accessed September 15, 2011. https://sites.google.com/site/acrsta/ids_meta-analysis. [p. 26]

CSJ DVD (2004). *The Corpus of Spontaneous Japanese, offline documentation*. A collection of 17 PDF files in Japanese, bundled in the `/DOC` directory of the DVD "Volume 1". [pp. 26, 27, 28, 29, 30, 41, and 47]

CSJ Website (2012). *The Corpus of Spontaneous Japanese, online documentation*. Accessed January 15, 2012. http://www.ninjal.ac.jp/products-k/katsudo/seika/corpus/public/. [pp. 25 and 27]

Daland, R. and J. B. Pierrehumbert (2011). Learning diphone-based segmentation. *Cognitive Science 35*(1), 119–155. [pp. 20 and 26]

Dautriche, I. (2009). Modélisation des processus d'acquisition du langage par des méthodes statistiques. Master's thesis, Institut National des Sciences Appliquées, Toulouse. [pp. 21, 22, 26, 27, 28, 33, 34, 35, 36, 37, 38, 51, 54, 55, 56, 61, 64, 74, 75, 82, 94, 95, 98, 106, and 151]

de Leeuw, J. (1988). Convergence of the majorization method for multidimensional scaling. *Journal of Classification 5*, 163–180. [p. 117]

de Leeuw, J. and P. Mair (2009). Multidimensional scaling using majorization: SMACOF in R. *Journal of Statistical Software 31*(3), 1–30. Version 1.0-0. [pp. 116, 117, 118, 119, and 120]

Dell, F. (1985). *Les règles et les sons, introduction à la phonologie générative* (second ed.). Collection Savoir. Paris: Hermann. [p. 14]

Dillon, B., E. Dunbar, and W. Idsardi (2012). A single stage approach to learning phonological categories: insights from Inuktitut. *Cognitive Science*. To appear. [p. 22]

Dudoit, S. and J. Fridlyand (2002). A prediction-based resampling method for estimating the number of clusters in a dataset. *Genome Biology 3*(7), 1–21. [p. 125]

Duin, R. P. W. (1999). Compactness and complexity of pattern recognition problems. In C. Perneel (ed.), *Proceedings of the International Symposium on Pattern Recognition "In Memoriam Pierre Devijver"*, Brussels, pp. 124–128. [pp. 16, 53, 116, 125, and 153]

Duin, R. P. W. and E. Pękalska (2012). The dissimilarity space: bridging structural and statistical pattern recognition. *Pattern Recognition Letters 33*, 826–832. [p. 121]

Dunbar, E. (2009). Pitfalls of distributional allophone learning. Presentation given at the Montreal-Ottawa-Toronto Phonology Workshop. [p. 59]

Dunn, J. C. (1973). A fuzzy relative of the ISODATA process and its use in detecting well-separated

clusters. *Journal of Cybernetics 3*(3), 32–57. [p. 125]

Eifring, H. B. and R. Theil (2004). *Linguistics for Students of Asian and African Languages*. Institutt for østeuropeiske og orientalske studier. [p. 13]

Ellis, N. C. (2002). Frequency effects in language processing: a review with implications for theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition 24*, 143–188. [p. 19]

Elman, J. L. (1990). Finding structure in time. *Cognitive Sciences 14*, 179–211. [pp. 20 and 61]

Elsner, M., S. Goldwater, and J. Eisenstein (2012). Bootstrapping a unified model of lexical and phonetic acquisition. In *Proceedings of the Fiftieth Annual Meeting of the Association of Computational Linguistics*, Jeju Island. [p. 20]

Esmaelnejad, J., J. Habibi, and S. H. Yeganeh (2010). A novel method to find appropriate for ϵ for DBSCAN. In N. T. Nguyen, M. T. Le, and J. Świątek (eds.), *Proceedings of the Second International Conference on Intelligent Information and Database Systems: Part I*, Berlin, Heidelberg, pp. 93–102. Springer-Verlag. [pp. 134 and 135]

Esposito, F., D. Malerba, V. Tamma, and H.-H. Bock (2000). Classical resemblance measures. In H.-H. Bock and E. Diday (eds.), *Analysis of Symbolic Data. Exploratory methods for extracting statistical information from complex data.*, Chapter 8. Heidelberg: Springer Verlag. [p. 64]

Ester, M., H.-P. Kriegel, J. Sander, and X. Xu (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, Portland, Oregon. [pp. 133, 134, and 135]

Everson, L. R. H. and W. T. Penzhorn (1988). Experimental comparison on several distance measures for speech processing applications. In *Proceedings of the Southern African Conference on Communications and Signal Processing*, pp. 12–17. [p. 54]

Faraway, J. J. (2005). *Linear Models with R*. Boca Raton, Florida: Chapman & Hall / CRC. [pp. 83 and 86]

Faraway, J. J. (2006). *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models*. Boca Raton, Florida: Chapman & Hall / CRC. [pp. 83, 85, 102, and 103]

Farnetani, E. (1997). Coarticulation and connected speech processes. In W. J. Hardcastle and J. Laver (eds.), *The Handbook of Phonetic Sciences*, Chapter 12, pp. 371–404. Blackwell Publishers. [p. 16]

Ferrer-i-Cancho, R. and R. V. Solé (2001). Two regimes in the frequency of words and the origins of complex lexicons: Zipf's law revisited. *Journal of Quantitative Linguistics 8*(3), 165–173. [p. 32]

Ferrer-i-Cancho, R. and R. V. Solé (2003). Least effort and the origins of scaling in human language. *Proceedings of the National Academy of Sciences of the United States of America (PNAS) 100*(3), 788–791. [p. 32]

Floyd, R. W. (1962). Algorithm 97: shortest path. *Communications of the ACM 5*(6). [p. 98]

Forgy, E. W. (1965). Cluster analysis of multivariate data: efficiency vs. interpretability of classifications. *Biometrics 21*, 768–769. [p. 142]

Gale, W. A. and K. Church (1994). What's wrong with adding one? In N. Oostdijk and P. de Haan (eds.), *Corpus-Based Research into Language: In honour of Jan Aarts*, Amsterdam, pp. 189–200. [p. 60]

Gale, W. A. and G. Sampson (1995). Good–turing frequency estimation without tears. *Journal of Quantitative Linguistics 2*, 217–237. [pp. 60 and 157]

Gambell, T. and C. Yang (2004). Statistics learning and universal grammar: modeling word segmentation. In *Proceedings of the Twentieth International Conference on Computational Linguistics*, pp. 49–52. [pp. 20, 26, and 153]

Gelman, A. and J. Hill (2007). *Data analysis using regression and multilevel/hierarchical models*.

Analytical methods for social research. New York: Cambridge University Press. [p. 103]

Gildea, D. and D. Jurafsky (1996). Learning bias and phonological-rule induction. *Computational Linguistics 22*, 497–530. [pp. 19, 21, 24, and 154]

Giorgino, T. (2009). Computing and visualizing dynamic time warping alignments in `R`: the `dtw` package. *Journal of Statistical Software 31*(7), 1–24. [p. 54]

Godfrey, P., R. Shipley, and J. Gryz (2007). Algorithms and analyses for maximal vector computation. *The International Journal on Very Large Data Bases 16*(1), 5–28. [p. 76]

Goldberger, J. and H. Aronowitz (2005). A distance measure between GMMs based on the unscented transform and its application to speaker recognition. In *Proceedings of the 9th European Conference on Speech Communication and Technology*, pp. 1985–1989. [pp. 37, 55, and 56]

Goldsmith, J. and A. Xanthos (2009). Learning phonological categories. *Language 85*(1), 4–38. [pp. 21 and 24]

Goldsmith, J. A., J. Riggle, and A. C. L. Yu (eds.) (2011). *The Handbook of Phonological Theory* (second ed.). Wiley–Blackwell. [p. 14]

Goldwater, S., T. L. Griffiths, and M. Johnson (2009). A Bayesian framework for word segmentation: exploring the effects of context. *Cognition 112*(1), 21–54. [pp. 20 and 61]

Gorawski, M. and R. Malczok (2006a). Towards automatic ϵ calculation in density-based clustering. In Y. Manolopoulos, J. Pokorný, and T. Sellis (eds.), *Advances in Databases and Information Systems*, Volume 4152 of *Lecture Notes in Computer Science*, pp. 313–328. Berlin, Heidelberg: Springer. [p. 135]

Gorawski, M. and R. Malczok (2006b). AEC algorithm: a heuristic approach to calculating density-based clustering ϵ parameter. In T. Yakhno and E. Neuhold (eds.), *Advances in Information Systems*, Volume 4243 of *Lecture Notes in Computer Science*, pp. 90–99. Berlin, Heidelberg: Springer. [p. 135]

Goutte, C. and E. Gaussier (2005). A probabilistic interpretation of precision, recall and F-score, with implication for evaluation. In *Proceedings of the 27th European Conference on Information Retrieval*, pp. 345–359. [p. 87]

Graham, R. L., D. E. Knuth, and O. Patashnik (1988). *Concrete Mathematics*. Reading, Massachusetts: Addison–Wesley. Sixth printing, with corrections. [pp. 141 and 142]

Gray, A. and J. Markel (1976). Distance measures for speech processing. *IEEE Transactions on Acoustics, Speech, and Signal Processing 24*(5). [p. 54]

Groenen, P. J. F. and W. J. Heiser (1996). The tunneling method for global optimization in multidimensional scaling. *Psychometrika 61*(3), 529–550. [p. 120]

Groenen, P. J. F. and W. J. Heiser (2000). "Optimization transfer using surrogate objective functions:" discussion. *Journal of Computational and Graphical Statistics 9*(1), 44–48. [p. 117]

Groenen, P. J. F. and M. van de Velden (2004). Multidimensional scaling. Technical Report EI 2004-15, Econometric Institute, Erasmus University Rotterdam. [pp. 116, 120, and 152]

Grønnum, N. and H. Basbøll (2003). Two-phased stød vowels—a cognitive reality? *Phonum 9*, 33–36. Umeå University, Department of Philosophy and Linguistics. [p. 56]

Ha, L. Q., E. I. Sicilia-Garcia, J. Ming, and F. J. Smith (2002). Extension of Zipf's law to words and phrases. In *Proceedings of the Nineteenth International Conference on Computational Linguistics*, pp. 315–320. [p. 32]

Halkidi, M., Y. Batistakis, and M. Vazirgiannis (2001). On clustering validation techniques. *Intelligent Information Systems Journal 17*(2–3), 107–145. [p. 125]

Halkidi, M. and M. Vazirgiannis (2001). Clustering validity assessment: finding the optimal partitioning of a data set. In *Proceedings of the IEEE International Conference on Data Mining*, pp. 187–194. [p. 125]

Harris, Z. S. (1951). *Methods in Structural Linguistics*. Chicago: University of Chicago Press.

[pp. 21 and 58]

Harris, Z. S. (1955). From phoneme to morpheme. *Language 31*, 192–222. [p. 21]

Hartigan, J. A. and M. A. Wong (1979). A K-means clustering algorithm. *Applied Statistics 28*, 100–108. [p. 142]

Hayes, B. (2004). Phonological acquisition in optimality theory: the early stages. In R. Kager, J. Pater, and W. Zonneveld (eds.), *Fixing Priorities: Constraints in Phonological Acquisition*. Cambridge: Cambridge University Press. [p. 21]

Hermansky, H. (1999). Mel cepstrum, deltas, double-deltas, ... what else is new? In *Proceedings of the Workshop on Robust Methods for Speech Recognition in Adverse Conditions*, Tampere. [p. 35]

Herrnstein, R. J., D. H. Loveland, and C. Cable (1976). Natural concepts in pigeons. *Journal of Experimental Psychology: Animal Behavior Processes 2*(4), 285–302. Inaccessible, cited by Martin et al. (2009). [pp. 68, 69, 101, and 106]

Hershey, J. R. and P. A. Olsen (2007). Approximating the Kullback–Leibler divergence between Gaussian mixture models. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Volume 4, pp. 317–320. [p. 55]

Hinds, J. (1990). *Japanese*. Croom Helm descriptive grammars. Croom Helm. [pp. 28, 29, 31, and 41]

Hockett, C. (1955). A manual of phonology. *International Journal of American Linguistics 21*(4). [p. 63]

Hodges, A. (2007). *One to Nine: the Inner Life of Numbers*. New York & London: W. W. Norton & Company. [p. 23]

Horan, C. B. (1969). Multidimensional scaling: Combining observations when individuals have different perceptual structures. *Psychometrika 34*, 139–65. [pp. 118 and 119]

Hsu, C.-W., C.-C. Chang, and C.-J. Lin (2010). A practical guide to support vector classification. Technical report, National Taiwan University. [p. 135]

Hubert, L. J. and P. Arabie (1985). Comparing partitions. *Journal of Classification 2*, 193–218. [p. 125]

Hubert, L. J., P. Arabie, and J. Meulman (2001). *Combinatorial data analysis: optimization by dynamic programming*. Philadelphia: SIAM. [p. 141]

Huijbregts, M., M. McLaren, and D. van Leeuwen (2011). Unsupervised acoustic sub-word unit detection for query-by-example spoken term detection. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4436–4439. [p. 20]

Hájek, A. (2012). Interpretations of probability. In E. N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2012 ed.). [p. 19]

Höppner, F. (2009). How much *true* structure has been discovered? In P. Perner (ed.), *Machine Learning and Data Mining in Pattern Recognition*, Volume 5632 of *Lecture Notes in Computer Science*, pp. 385–397. Springer Berlin / Heidelberg. [pp. 125 and 132]

International Phonetic Association (1999). Phonetic description and the IPA chart. In *Handbook of the International Phonetic Association: a guide to the use of the international phonetic alphabet*. Cambridge University Press. [pp. 14 and 32]

Ito, J. and A. Mester (1995). Japanese phonology. In J. Goldsmith (ed.), *The Handbook of Phonological Theory*, pp. 817–838. Blackwell. [pp. 28, 31, 41, and 45]

Jain, A. K. (2008). Data clustering: 50 years beyond k-means. In *Proceedings of the 2008 European Conference on Machine Learning and Knowledge Discovery in Databases: Part I*, Berlin, Heidelberg, pp. 3–4. Springer-Verlag. [p. 125]

Jain, A. K., M. N. Murty, and P. J. Flynn (1999). Data clustering: a review. *ACM Computing Surveys 31*(3), 264–323. [pp. 125 and 132]

Johnson, E. K. and P. W. Jusczyk (2001). Word segmentation by 8-month olds: when speech cues

count more than statistics. *Journal of Memory and Language 44*, 548–567. [p. 20]

Johnson, M. (1984). A discovery procedure for certain phonological rules. In *Proceedings of the Tenth International Conference on Computational Linguistics and Twenty-Second Annual Meeting of the Association for Computational Linguistics*, pp. 344–347. [p. 21]

Johnson, M. (2008a). Using adaptor grammars to identify synergies in the unsupervised acquisition of linguistic structure. In *Proceedings of Forty-Sixth Annual Meeting of the Association for Computational Linguistics*, Columbus, Ohio, pp. 398–406. Association for Computational Linguistics. [p. 154]

Johnson, M. (2008b). Unsupervised word segmentation for Sesotho using adaptor grammars. In *Proceedings of the Tenth Meeting of ACL Special Interest Group on Computational Morphology and Phonology*, pp. 20–27. [pp. 20 and 154]

Johnson, M., T. L. Griffiths, and S. Goldwater (2007). Adaptor grammars: a framework for specifying compositional nonparametric bayesian models. In B. Schoelkopf, J. Platt, and T. Hoffman (eds.), *Advances in Neural Information Processing Systems*, Volume 19. The MIT Press. [p. 154]

Jones, E., T. Oliphant, P. Peterson, et al. (2001). SciPy: open source scientific tools for Python. http://www.scipy.org/. [pp. 65 and 69]

Jurafksy, D. and J. H. Martin (2009). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Processing* (second ed.). Upper Saddle River, New Jersey: Prentice Hall. [pp. 16, 22, 35, 36, and 102]

Jusczyk, P. W. and J. Bertoncini (1988). Viewing the development of speech perception as an innately guided process. *Language and Speech 31*, 217–238. [pp. 19, 24, and 154]

Kager, R. (1999). *Optimality Theory*. Cambridge Textbooks in Linguistics. Cambridge: Cambridge University Press. [pp. 14 and 21]

Kailing, K., H.-P. Kriegel, and P. Krögers (2004a). Density-connected subspace clustering for high-dimensional data. In *Proceedings of the SIAM International Conference on Data Mining*, pp. 246–257. [p. 135]

Kailing, K., H.-P. Kriegel, A. Pryakhin, and M. Schubert (2004b). Clustering multi-represented objects with noise. In *Proceedings of the Eighth Pacific-Asia Conference on Knowledge Discovery and Data Mining*, Sydney, pp. 394–403. [pp. 133 and 135]

Kaplan, R. M. and M. Kay (1994). Regular models of phonological rule systems. *Computational Linguistics 20*(3), 331–378. [pp. 21, 22, and 24]

Kauffman, L. and P. J. Rousseeuw (1990). *Finding groups in data: an introduction to cluster analysis*. New York: Wiley. [p. 141]

Kazanina, N., C. Phillips, and W. Idsardi (2006). The influence of meaning on the perception of speech sounds. *Proceedings of the National Academy of Sciences of the United States of America (PNAS) 103*(30), 11381–11386. [p. 14]

Kewley-Port, D., T. Z. Burkle, and J. H. Lee (2007). Contribution of consonant versus vowel information to sentence intelligibility for young normal-hearing and elderly hearing-impaired listeners. *Journal of the Acoustical Society of America 122*, 2365–2375. [p. 56]

Kim, M. and R. S. Ramakrishna (2005). New indices from cluster validity assessment. *Pattern Recognition Letters 26*, 2353–2363. [pp. 100 and 125]

Kinda-Ichi, H. and H. Maës (1978). Phonologie du japonais standard. Volume V of *Travaux de linguistique japonaise*. Université de Paris 7. [pp. 28, 29, 30, 41, and 45]

Klein, D. and C. Manning (2003). MaxEnt models, conditional estimation, and optimization without the magic. Tutorial given at ACL 2003. [p. 102]

Koskenniemi, K. (1983). Two-level morphology: a general computational model of word-form recognition and production. Technical Report #11, Department of General Linguistics, Univer-

sity of Helsinki. [p. 24]

Kuhl, P. K. (2004). Early language acquisition: Cracking the speech code. *Nature Reviews Neuroscience 5*(11), 831–843. [pp. 15, 18, 19, 22, 24, 26, and 61]

Kuhn, H. W. (1955). The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly 2*, 83–97. [p. 56]

Kullback, S. and R. Leibler (1951). On information and sufficiency. *Annals of Mathematical Statistics 22*, 76–86. [p. 54]

Le Calvez, R. (2007). *Approche computationnelle de l'acquisition précoce des phonèmes*. Ph. D. thesis, Université Pierre et Marie Curie, Paris. [pp. 21, 22, 24, 34, 51, 53, 58, 59, 61, 64, 75, 82, 95, and 98]

Le Calvez, R., S. Peperkamp, and E. Dupoux (2007). Bottom-up learning of phonemes: A computational study. In *Proceedings of the Second European Cognitive Science Conference*, pp. 167–172. [pp. 21, 22, 34, 51, 53, 61, 75, 82, 95, 98, and 151]

Legány, C., S. Juhász, and A. Babos (2006). Cluster validity measurement techniques. In *Proceedings of the Fifth International Conference on Artificial Intelligence, Knowledge Engineering and Data Bases*, Stevens Point, Wisconsin, pp. 388–393. World Scientific and Engineering Academy and Society. [p. 125]

Lin, J. (1991). Divergence measures based on the Shannon entropy. *IEEE Transactions of Information Theory 37*(1), 145–151. [p. 59]

Lloyd, S. P. (1957). Least squares quantization in PCM. Technical report, Bell Laboratories. Published in 1982 in *IEEE Transactions on Information Theory 28*, 128-137. [p. 142]

Lowe, C. R., C. J. Roberts, and S. Lloyd (1971). Malformations of central nervous system and softness of local water supplies. *British Medical Journal 2*, 357–361. [p. 103]

Lupşa, D. A. (2005). Unsupervised single-link hierarchical clustering. *Studia Universitatis Babes-Bolyai, Series Informatica L*(2), 11–22. [p. 143]

Lyons, J. (1968). *Introduction to Theoretical Linguistics* (1995 ed.). Cambridge University Press. [pp. 15, 16, 17, 51, and 63]

Machata, M. and Z. Jelaska (2006). Prototypicality and the concept of phoneme. *Glossos 6*. [p. 15]

Macneilage, P. F. (1997). Acquisition of speech. In W. J. Hardcastle and J. Laver (eds.), *The Handbook of Phonetic Sciences*, Chapter 10, pp. 301–332. Blackwell Publishers. [p. 18]

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In L. M. Le Cam and J. Neyma (eds.), *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Volume 1, Berkeley, California, pp. 281–297. University of California Press. [p. 142]

Maekawa, K. (2003). Corpus of spontaneous Japanese: Its design and evaluation. In *Proceedings of the ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, Tokyo. [pp. 25 and 153]

Maekawa, K., H. Koiso, S. Furui, and H. Isahara (2000). Spontaneous speech corpus of Japanese. In *Proceedings of the Second International Conference of Language Resources and Evaluation*, Athens, pp. 947–952. [pp. 25 and 153]

Magri, G. (2012). The error-driven ranking model of the acquisition of phonotactics. In *Proceedings of the Sixth Workshop on Psychocomputational Models of Human Language Acquisition*, Portland, Oregon. [p. 21]

Mak, B. and E. Barnard (1996). Phone clustering using the Bhattacharyya distance. In *Proceedings of the Fourth International Conference on Spoken Language Processing*, Philadelphia, Pennsylvania. International Speech Communication Association. [p. 54]

Makhoul, J. and R. Schwartz (1995). State of the art in continuous speech recognition. *Proceedings of the National Academy of Sciences of the United States of America (PNAS) 92*, 9956–9963. [p. 35]

Manning, C. D., P. Raghavan, and H. Schütze (2008). *Introduction to Information Retrieval*. Cambridge University Press. [pp. 105, 125, 128, 136, and 143]

Marr, D. (1982). *Vision: a Computational Investigation into the Human Representation and Processing of Visual Information*. San Francisco: W. H. Freeman & Company. [pp. 23 and 57]

Martin, A., S. Peperkamp, and E. Dupoux (2009). Learning phonemes with a pseudo-lexicon. Submitted. [pp. 21, 22, 24, 26, 33, 34, 35, 39, 51, 61, 62, 63, 64, 66, 68, 74, 75, 82, 94, 95, 98, 106, 151, and 162]

Matsuda, S., M. Nakai, H. Shimodaira, and S. Sagayama (2000). Feature-dependent allophone clustering. In *Proceedings of the Sixth International Conference on Spoken Language Processing*, Volume 1, pp. 413–416. International Speech Communication Association. [p. 20]

Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta, Protein Structure 405*(2), 442–451. [p. 88]

McCallum, A., K. Nigam, and L. H. Ungar (2000). Efficient clustering of high-dimensional data sets with application to reference matching. In *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, New York, New York, pp. 169–178. ACM. [pp. 33 and 66]

McCullagh, P. and J. A. Nelder (1989). *Generalized Linear Models* (second ed.). London: Chapman & Hall / CRC. [pp. 103 and 104]

MDS(X) User Manual (1981). *The MDS(X) User Manual*. Edinburgh. [pp. 118 and 119]

Meila, M. (2007). Comparing clusterings: an information-based distance. *Journal of Multivariate Analysis 98*, 873–895. [pp. 125 and 133]

Mermelstein, P. (1976). Distance measures for speech recognition, psychological and instrumental. In *Proceedings of the Joint Workshop on Pattern Recognition and Artificial Intelligence*, pp. 374–388. [p. 54]

Miller, R. A. (1967). *The Japanese Language*. History and Structure of Languages. Midway reprints. Chicago & London: The University of Chicago Press. [pp. 15, 28, 29, 41, 45, and 58]

Milligan, G. W. and M. C. Cooper (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika 50*(2), 159–179. [p. 125]

Mirkin, B. G. (2005). *Clustering for Data Mining: A Data Recovery Approach*. Boca Raton, Florida: Chapman & Hall / CRC. [pp. 58, 64, 65, 125, 132, 142, and 143]

Mohri, M. and B. Roark (2005). Structural zeros versus sampling zeros. Technical Report #CSE-05-003, Computer Science & Electrical Engineering, Oregon Health & Science University. [p. 61]

Mount, J. (2011). The equivalence of logistic regression and maximum entropy models. Accessed April 15, 2012. http://www.win-vector.com/dfiles/LogisticRegressionMaxEnt.pdf. [p. 102]

Munkres, J. (1957). Algorithms for the assignment and transportation problems. *Journal of the Society for Industrial and Applied Mathematics 5*(1), 32–38. [pp. 56 and 159]

Ng, R. T. and J. Han (1994). Efficient and effective clustering methods for spatial data mining. In *Proceedings of the Twentieth International Conference on Very Large Data Bases*, Santiago, pp. 144–155. [p. 133]

Norris, D. (2005). How do computational models help us build better theories? In A. Cutler (ed.), *Twenty-First Century Psycholinguistics: Four Cornerstones*, Chapter 20. [p. 23]

Ohala, J. J. (1997). The relation between phonetics and phonology. In W. J. Hardcastle and J. Laver (eds.), *The Handbook of Phonetic Sciences*, Chapter 22, pp. 674–694. Blackwell Publishers. [p. 24]

Okada, H. (1999). Japanese. In *Handbook of the International Phonetic Association: A guide to the usage of the International Phonetic Alphabet*, pp. 117–119. Cambridge: Cambridge University Press. [pp. 28, 29, 30, 32, and 41]

Olivier, D. C. (1968). *Stochastic Grammars and Language Acquisition Mechanisms*. Ph. D. thesis,

Harvard University. Inaccessible, cited by Brent (1999). [pp. 20 and 61]

Otsu, Y. (1980). Some aspects of *rendaku* in Japanese and related problems. In Y. Otsu and A. Farmer (eds.), *Theoretical Issues in Japanese Linguistics, MIT Working Papers in Linguistics* (third ed.), Volume 2, pp. 207–227. [p. 28]

Papadias, D., Y. Tao, G. Fu, and B. Seeger (2005). Progressive skyline computation in database systems. *ACM Transactions on Database Systems 30*(1), 41–82. [p. 76]

Pardo, C. E. and P. C. DelCampo (2007). Combinacion de metodos factoriales y de analisis de conglomerados en `R`: el paquete `FactoClass`. *Revista Colombiana de Estadistica 30*(2), 231–245. Version 1.0.8. [p. 142]

Patterson, D. E. and M. R. Berthold (2001). Clustering in parallel universes. In *IEEE Conference on Systems, Man and Cybernetics*. IEEE Press. [p. 121]

Pearl, L., S. Goldwater, and M. Steyvers (2010). Online learning mechanisms for Bayesian models of word segmentation. *Research on Language and Computation 8*(2), 107–132. [pp. 20 and 61]

Pękalska, E. and R. P. W. Duin (2005). *The Dissimilarity Representation for Pattern Recognition. Foundations and Applications.* Singapore: World Scientific. [pp. 23 and 117]

Pękalska, E., D. M. J. Tax, and R. P. W. Duin (2003). One-class LP classifiers for dissimilarity representations. In S. Becker, S. Thrun, and K. Obermayer (eds.), *Advances in Neural Information Processing Systems*, Volume 15, Cambridge, pp. 761–768. MIT Press. [pp. 16, 53, 116, 125, and 153]

Peperkamp, S. (2003). Phonological acquisition: recent attainments and new challenges. *Language and Speech 46*(2–3), 87–113. [pp. 18, 19, and 20]

Peperkamp, S. and E. Dupoux (2002). Coping with phonological variation in early lexical acquisition. In I. Lasser (ed.), *The Process of Language Acquisition*, pp. 359–385. Frankfurt: Peter Lang. [pp. 21, 24, 58, and 61]

Peperkamp, S., R. Le Calvez, J.-P. Nadal, and E. Dupoux (2006). The acquisition of allophonic rules: statistical learning with linguistic constraints. *Cognition 101*(3), B31–B41. [pp. 7, 11, 12, 21, 22, 24, 26, 33, 34, 38, 39, 51, 52, 53, 58, 59, 60, 61, 64, 65, 66, 68, 74, 75, 82, 83, 84, 88, 94, 95, 97, 98, 99, 100, 106, 111, 114, 115, 127, 137, 150, 151, 152, and 153]

Plunkett, K. (1997). Theories of early language acquisition. *Trends in Cognitive Sciences 1*(4), 146–153. [pp. 18, 19, 24, and 154]

Prince, A. and P. Smolensky (1993). Optimality theory: constraint interaction in generative grammar. ROA version, #ROA-537, Rutgers, The State University of New Jersey. [p. 21]

PsychoCompLA. Psychocomputational models of human language acquisition. http://www.colag.cs.hunter.cuny.edu/psychocomp/. [p. 22]

R Development Core Team (2010). *R: a Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. Version 2.12.1. http://www.R-project.org/. [pp. 85, 117, and 142]

Ramus, F., S. Peperkamp, A. Christophe, C. Jacquemot, S. Kouider, and E. Dupoux (2010). A psycholinguistic perspective on the acquisition of phonology. In C. Fougeron, B. Kühnert, M. D'Imperio, and N. Vallé (eds.), *Laboratory Phonology 10*, pp. 311–340. Berlin: Mouton de Gruyter. [p. 21]

Rau, D. V., H.-H. A. Chang, and E. E. Tarone (2009). Think or sink: Chinese learners' acquisition of the english voiceless interdental fricative. *Language Learning 59*(3), 581–621. [p. 14]

Reichart, R. and A. Rappoport (2009). The NVI clustering evaluation measure. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*. [pp. 125, 132, 133, 135, and 136]

Reynolds, D. A. (2009). Gaussian mixture models. In *Encyclopedia of Biometrics*, pp. 659–663. [p. 37]

Rodríguez, G. (2007). Lecture notes on generalized linear models. Accessed April 15, 2012. http://data.princeton.edu/wws509/notes/. [p. 103]

Rosenberg, A. and J. Hirschberg (2007). V-measure: a conditional entropy-based external cluster evaluation measure. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. [p. 133]

Rosman, G., A. M. Bronstein, M. M. Bronstein, A. Sidi, and R. Kimmel (2008). Fast multidimensional scaling using vector extrapolation. Technical Report CIS-2008-01, Technion, Computer Science Department. [p. 117]

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics 20*, 53–65. [pp. 125 and 126]

Rousseeuw, P. J. and C. Croux (1992). Explicit scale estimators with high breakdown point. In Y. Dodge (ed.), *L_1-Statistical Analysis and Related Methods*, pp. 77–92. Amsterdam: North Holland. [p. 57]

Rousseeuw, P. J. and C. Croux (1993). Alternatives to the median absolute deviation. *Journal of the American Statistical Association 88*(424), 1273–1283. [p. 57]

Rubinov, A., N. Soukhoroukova, and J. Ugon (2006). Classes and clusters in data analysis. *European Journal of Operational Research 173*, 849–865. [p. 125]

Ruiz, C., M. Spiliopoulou, and E. Menasalvas (2007). C-DBSCAN: Density-based clustering with constraints. In A. An, J. Stefanowski, S. Ramanna, C. Butz, W. Pedrycz, and G. Wang (eds.), *Rough Sets, Fuzzy Sets, Data Mining and Granular Computing*, Volume 4482 of *Lecture Notes in Computer Science*, pp. 216–223. Berlin, Heidelberg: Springer. [p. 135]

Rytting, C. A., C. Brew, and E. Fosler-Lussier (2010). Segmenting words from natural speech: subsegmental variation in segmental cues. *Journal of Child Language 37*, 513–543. [p. 20]

Saffran, J. R. (2002). Constraints on statistical language learning. *Journal of Memory and Language 47*, 172–196. [pp. 19 and 24]

Saffran, J. R., R. N. Aslin, and E. L. Newport (1996). Statistical learning by 8-month-old infants. *Science 274*, 1926–1928. [pp. 19 and 24]

Sakoe, H. and S. Chiba (1978). Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing 26*(1), 43–49. [p. 54]

Sander, J., M. Ester, H.-P. Kriegel, and X. Xu (1998). Density-based clustering in spatial databases: the algorithm GDBSCAN and its applications. *Data Mining and Knowledge Discovery 2*(2), 169–194. [pp. 133 and 135]

Schwarz, P., P. Matějka, and J. Černocký (2004). Towards lower error rates in phoneme recognition. In P. Sojka, I. Kopecek, and K. Pala (eds.), *Text, Speech and Dialogue*, Volume 3206 of *Lecture Notes in Computer Science*, pp. 465–472. Berlin, Heidelberg: Springer. [p. 20]

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal 27*, 379–423, 623–656. [p. 13]

Shepard, R. N. (1962a). The analysis of proximities: multidimensional scaling with an unknown distance function. I. *Psychometrika 27*(2), 125–140. [p. 117]

Shepard, R. N. (1962b). The analysis of proximities: multidimensional scaling with an unknown distance function. II. *Psychometrika 27*(3), 219–246. [p. 117]

Sing, T., O. Sander, N. Beerenwinkel, and T. Lengauer (2005). `ROCR`: visualizing classifier performance in `R`. *Bioinformatics 21*(20), 3940–3941. http://rocr.bioinf.mpi-sb.mpg.de/. [p. 87]

Steinley, D. (2007). Validating clusters with the lower bound for sum-of-squares error. *Psychometrika 72*(1), 93–106. [p. 141]

Strehl, A., J. Ghosh, and R. J. Mooney (2000). Impact of similarity measures on web-page clustering. In *Proceedings of the AAAI Workshop on Artificial Intelligence for Web Search*, Austin, Texas, pp. 58–64. AAAI/MIT Press. [p. 125]

Stylianou, Y. (2008). Voice transformation. In J. Benesty, M. M. Sondhi, and Y. Huand (eds.), *Springer handbook of speech processing*, Chapter 24, pp. 489–503. Springer Verlag. [p. 37]

Surendran, D. and P. Niyogi (2006). Quantifying the functional load of phonemic oppositions, distinctive features, and suprasegmentals. In O. Nedergaard Thomsen (ed.), *Competing Models of Language Change: Evolution and Beyond*, Amsterdam & Philadelphia. John Benjamins. [p. 63]

Takane, Y., F. W. Young, and J. de Leeuw (1977). Nonmetric individual differences multidimensional scaling: an alternating least squares method with optimal scaling features. *Psychometrika 42*(1), 7–67. [p. 120]

Tambovtsev, Y. and C. Martindale (2007). Phoneme frequencies follow a Yule distribution. *SKASE Journal of Theoretical Linguistics 4*(2), 1–11. [pp. 14 and 32]

Tesar, B. and P. Smolensky (1996). Learnability in optimality theory. Technical Report #JHU-CogSci-96-3, Johns Hopkins University. [p. 21]

Thiessen, E. D. and J. R. Saffran (2003). When cues collide: use of stress and statistical cues to word boundaries by 7- to 9-month old infants. *Developmental Psychology 39*(4), 706–716. [p. 20]

Tibshirani, R., G. Walther, and T. Hastie (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society B 63*(2), 441–423. [p. 125]

Tobin, Y. (1997). *Phonology as Human Behavior: Theoretical Implications and Clinical Applications*. Sound and Meaning: The Roman Jakobson Series in Linguistics and Poetics. Durham, North Carolina: Duke University Press. [p. 15]

Torgerson, W. S. (1952). Multidimensional scaling: I. Theory and method. *Psychometrika 17*, 401–419. [pp. 116 and 152]

Trubetzkoy, N. (1939). *Principles of Phonology (Grundzüge der Phonologie)* (sixth ed.). University of California Press. [p. 54]

Vakharia, A. J. and U. Wemmerlöv (1995). A comparative investigation of hierarchical clustering techniques and dissimilarity measures applied to the cell formation problem. *Journal of Operations Management 13*(2), 117–138. [p. 64]

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 of the United States of America (PNAS) 104*(33), 13273–13278. [p. 22]

van Os, B. J. (2000). *Dynamic programming for partitioning in multivariate data analysis*. Leiden University Press. [p. 141]

Vance, T. J. (1980). Comments on "Some aspects of *rendaku* in Japanese and related problems". In Y. Otsu and A. Farmer (eds.), *Theoretical Issues in Japanese Linguistics, MIT Working Papers in Linguistics* (third ed.), Volume 2, pp. 229–236. [p. 28]

Vance, T. J. (2008). *The Sounds of Japanese*. Cambridge University Press. [pp. 28, 29, 41, and 50]

Venables, W. N. and B. D. Ripley (2002). *Modern Applied Statistics with S* (4th ed.). New York: Springer. http://www.stats.ox.ac.uk/pub/MASS4. [pp. 103, 117, 120, and 127]

Venkataraman, A. (2001). A statistical model for word discovery in transcribed speech. *Computational Linguistics 27*(3), 351–372. [pp. 20 and 61]

Wade, G. (1994). *Signal coding and processing* (second ed.). Cambridge University Press. [p. 15]

Waibel, A., T. Hanazawa, G. Hinton, K. Shikano, and K. J. Lang (1989). Phoneme recognition using time-delay neural networks. *IEEE Transactions on Acoustics, Speech, and Signal Processing 37*(3), 328–339. [p. 20]

Wang, C.-C., Q.-L. Ding, H. Tao, and H. Li (2012). Comment on "Phonemic diversity supports a serial founder effect model of language expansion from Africa". *Science 335*(6069), 657. [p. 14]

Wang, X. and H. J. Hamilton (2003). DBRS: a density-based spatial clustering method with random sampling. Technical Report #CS-2003-13, Department of Computer Science, University of Regina. [p. 135]

Wang, X., C. Rostoker, and H. J. Hamilton (2004). Density-based spatial clustering in the presence of obstacles and facilitators. Technical Report #CS-2004-9, Department of Computer Science, University of Regina. [p. 135]

Ward, Jr., J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association 58*(301), 236–244. [p. 143]

Warshall, S. (1962). A theorem on Boolean matrices. *Journal of the ACM 9*(1), 11–12. [p. 98]

Weingessel, A., E. Dimitriadou, and S. Dolnicar (1999). An examination of indexes for determining the number of clusters in binary data sets. Working Paper #29, Vienna University of Economics. [p. 125]

Wells, J. C. (1999). British English pronunciation preferences: a changing scene. *Journal of the International Phonetic Association 29*, 33–50. [p. 31]

Werker, J. F. and R. C. Tees (1999). Influences on infant speech processing: toward a new synthesis. *Annual Review of Psychology 50*, 509–535. [p. 18]

Wikipedia (2012). Japanese phonology. In *Wikipedia, the free encyclopedia*. Accessed March 15, 2012. http://en.wikipedia.org/w/index.php?title=Japanese_phonology&oldid=482078373. [pp. 28, 29, 32, 41, and 45]

Wintner, S. (2010). Formal language theory. In A. Clark, C. Fox, and S. Lappin (eds.), *The Handbook of Computational Linguistics and Natural Language Processing*, Chapter 1, pp. 9–42. Wiley-Blackwell. [p. 22]

Wiswedel, B., F. Hoeppner, and M. R. Berthold (2010). Learning in parallel universes. *Data Mining and Knowledge Discovery 21*, 130–150. [p. 121]

Woodland, P. C. (2001). Speaker adaptation for continuous density HMMs: a review. In *Proceedings of the ISCA Tutorial and Research Workshop on Adaptation Methods for Speech Recognition*, Sophia Antipolis, pp. 11–19. [p. 19]

Young, S., G. Evermann, M. Gales, T. Hain, D. Kershaw, X. A. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland (2006). The HTK Book (for HTK version 3.4). Cambridge University Engineering Department. [pp. 35, 36, 37, 38, 47, and 56]

Yu, K. M. (2010). Representational maps from the speech signal to phonological categories: a case study with lexical tones. *UCLA Working Papers in Linguistics 15*(5), 1–30. [p. 22]

Zipf, G. K. (1935). *The Psychobiology of Language*. The International Library of Psychology. London: Routledge. 2002 reprint. [p. 32]