UCLA Working Papers in Linguistics, Papers in Mathematical Linguistics 1
Volume 15, Article 5: 1–30, 2010

Representational Maps from the Speech Signal to Phonological Categories: a Case Study with Lexical Tones

Kristine M. Yu

As the initial step in studying the acquisition of phonological categories from the speech signal, we describe representational issues for the target of learning, a probabilistic distribution of phonological categories over a phonetic parameter space. Our model system of study is cross-linguistic lexical tonal phonemes in tonal languages. We focus on two representational issues: temporal resolution of the extracted phonetic parameters and static and dynamic parameterizations of the speech signal. In a human perception study and exploratory computational modeling, we find that coarse sampling of absolute f0 and f0 velocity is sufficient for near-partitions of the phonetic parameter space for single-speaker tonal spaces in a range of tone languages.

Keywords phonetics, phonology, tone, learnability

Introduction

The phonetic realization of linguistic tone is widely believed to be simple and limited to a single dimension of fundamental frequency (the physical correlate of the auditory percept of pitch).

> ...tones typically involve a single primary acoustic dimension, namely, f0. This contrasts with the multiple acoustic dimensions such as formants or spectral peaks required for characterizing vowels and consonants. The variability problem with tones is therefore at least limited to a single dimension... (Gauthier, Shi, and Xu 2007:82)

> ...tone presents few, if any articulatory difficulties vs. consonants (which all languages have). Second, tone is acoustically (hence perceptually?) simple, F_0, vs. consonants and vowels. (Hyman 2010:1)

In this paper, we show that the phonological representation of tone in terms of phonetic parameters may indeed be simple, but not necessarily in the way that has been described above. Even an entirely f0-based parameterization of tone can be highly multidimensional, since we may choose multiple ways to parametrize f0, e.g. with f0 height values and with f0 velocity values, and we may choose to sample these values arbitrarily densely in time. Here we show that: (i) both f0 height and f0 velocity are relevant parameters for a range of tone languages, even for the simplest level tone languages, and (ii) the relation between phonetic space and tonal categories may be simple, but in a way that may not be unique to tone: we show that coarse sampling of the relevant parameters suffices for good category separability—a near partition of the phonetic space—in a range of languages, and that humans can identify tones degraded to be coarsely sampled at a comparable level of accuracy to that for intact tones.

The work in this paper bears on defining the target of learning in the acquisition of lexical tonal categories from the speech signal, the initial step towards answering our larger research questions: (i) what the relation between phonetic spaces and tonal phonological categories is, i.e. how tones are phonetically realized, (ii) how that relation between the phonetic space and phonological categories could be learned, and (iii) how it is learned by L1 human learners. We frame our work broadly to scientifically explicate the universal structure in the phonetic parameter space across phonetically diverse tonal systems; we set up learning tonal categories as a model system for learning phonological categories to integrate the study of the acquisition of tone into the highly active research area of language acquisition in general. We take the broad perspective of Welmers (1973):

> In principle, the varieties and functions of tonal contrasts in language are of the same order as the varieties and functions of any other contrasts; the problems of tonal analysis are simply typical problems of linguistic analysis. (Welmers 1973:77)

Thus, we begin in §1 with preliminaries: we describe the learning problem in the context of phonological category acquisition, motivate the study of the target of learning (the map from the phonetic space to phonological categories), describe the larger research questions and methodological abstractions taken in the study and explicate our particular model system. In §2, we state the aspects of phonological representation, and more specifically, of tonal representation, that are the focus of this paper. These aspects are temporal resolution of the parameterized speech signal and static vs. dynamic properties of the speech signal. We end in §3 by briefly highlighting results from our own experiments and initial computational modeling work addressing these aspects.

1 Preliminaries

This paper investigates the learnability of lexical tonal phonological categories in tone languages. It is a preliminary step in the study of a much larger research question:

(Q0) *How do children acquire phonological categories from the speech signal?*

We address (Q0) using computational learning methods, like previous studies of phonological category learning, cf. de Boer and Kuhl (2003); Lin (2005); Toscano and McMurray (2010); Vallabha, McClelland, Pons, Werker, and Amano (2007), and moreover, we ground our modeling assumptions based on phonetic fieldwork and perception experiments we conducted.

While a complete answer to Q0 necessitates a battery of behavioral, physiological, production, and perceptual studies on infants from the womb to adulthood, particularly in the first years of life, our ability to probe infant knowledge of phonological categories and connect this knowledge to their language input is limited, cf. methodological approaches in Polka, Jusczyk, and Rvachew (1995); Werker, Shi, Desjardins, Pegg, and Polka (1998). Thus, we make the choice to generalize our study to *any learner* so that we can deploy mathematically-specified learners to learn from examples we have very fine control over. The advantage of focusing on computational approaches is that we can make a tight connection between the data that a learning machine gets (the domain of the learner, $\mathscr{D}$), how it went about the learning (the functional/algorithmic form of the learner, $\mathscr{A}$), and the target of learning (the codomain of the learner, $\mathscr{C}$). The challenge then is to also maintain a tight connection between the computational modeling and what we know about human learners.

Thus, we modify our original research question:

(Q0′) *How could a learner $\mathscr{A}: Data \to \mathscr{C}$ acquire lexical tonal categories from the speech signal in a way consistent with our knowledge about how humans do it?*

A key component in maintaining a tight connection between the computational modeling and human cognition is to have a clear picture of what the target of learning is (Dyson 2004; Minsky and Papert 1971). Thus, the goal of this paper is to define the codomain, $\mathscr{C}$, the target of learning in the acquisition of lexical tonal categories: we define what it means to have learned the lexical tonal categories of a tonal language; **we study the *learnability* of tonal spaces, conditioned on different representations of tonal examples, to understand how lexical tonal categories are defined**.

1.1 *The Target of Learning: the Phonetics-Phonology Map*

What does it mean to have learned the tones of a tone language, e.g. the tones of Mandarin: Tones 1-4, respectively, ˥ (high level), ˧˥ (rise), ˨˩˦ (fall-rise), ˥˩ (fall) (c.f. Fig. 1)? We assert that it means that the learner has learned a representational map:

(1) $$\mathscr{A}: Data \to \text{representational map}$$

and that this phonetics-phonology map is of the form:

(2) *Phonetics-Phonology:* $$\{\text{sequences of phonetic parameter vectors}\} \to \{\text{sets of phonological categories}\}$$

where the phonological categories are lexical tonal categories.

We show a familiar example of a well-studied phonetics-phonology map in Fig. 2, a vowel formant plot (Peterson and Barney 1952). This is a two-dimensional map

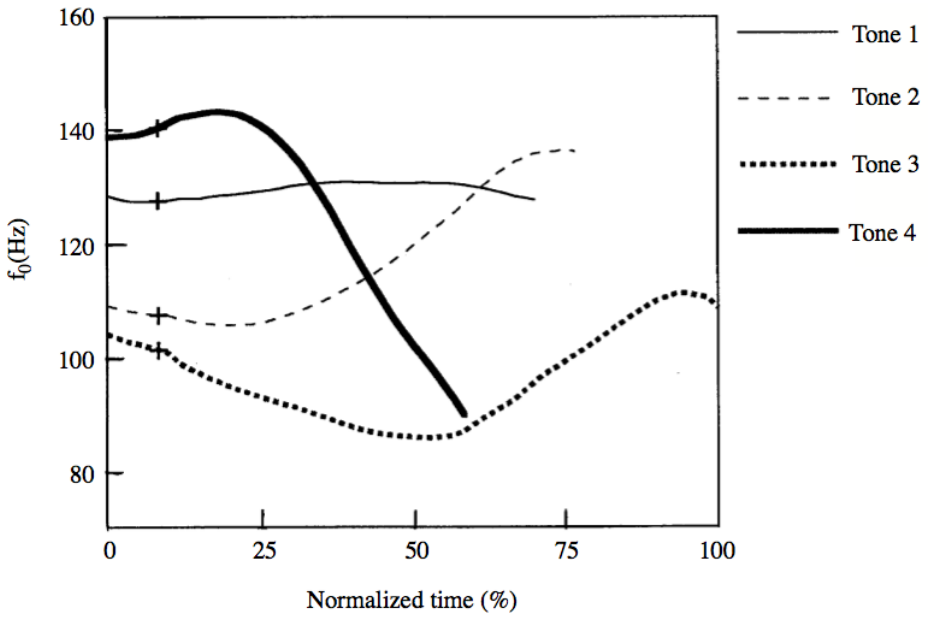


Figure 1: The tones of Mandarin (Xu 1997).

in $\langle F1_{SS}, F2_{SS}\rangle$ space (over the steady-state values of the first and second formants) which maps unit-length sequences of phonetic parameter vectors $\langle F1_{SS}, F2_{SS}\rangle$ to English vowel phonemes, cf. Table 1.

$\langle F1_{SS}, F2_{SS}\rangle$	English vowel phoneme	Note
$\langle 240, 2280\rangle$	{/i/}	Actual data point
$\langle 460, 1330\rangle$	{/ɝ/}	Actual data point
$\langle 475, 1220\rangle$	{/ʊ/}	Actual data point
$\langle 686, 1028\rangle$	{/ɑ, ɔ/}	Ambiguity
$\langle 400, 3500\rangle$	{/i/}	Not a data point
⋮	⋮	

Table 1: The representational map from steady state formant space to English vowel phonemic categories from Peterson and Barney (1952).

There are two things to note from Fig. 2 and Table 1 which are general properties of phonetics-phonology maps:

1. There are regions of $\langle F1_{SS}, F2_{SS}\rangle$ space where the same $\langle F1_{SS}, F2_{SS}\rangle$ point is mapped to multiple English vowel phonemes: regions where vowel ellipses overlap. This highlights that *ambiguity in phonetic-phonological maps implies a codomain of sets of phonological categories rather than of single phonological categories.*

2. The map is total within the vowel ellipses for $\langle F1_{SS}, F2_{SS}\rangle$ values, meaning that *all* $\langle F1_{SS}, F2_{SS}\rangle$ points included in the sets of $\langle F1_{SS}, F2_{SS}\rangle$ values bounded by the

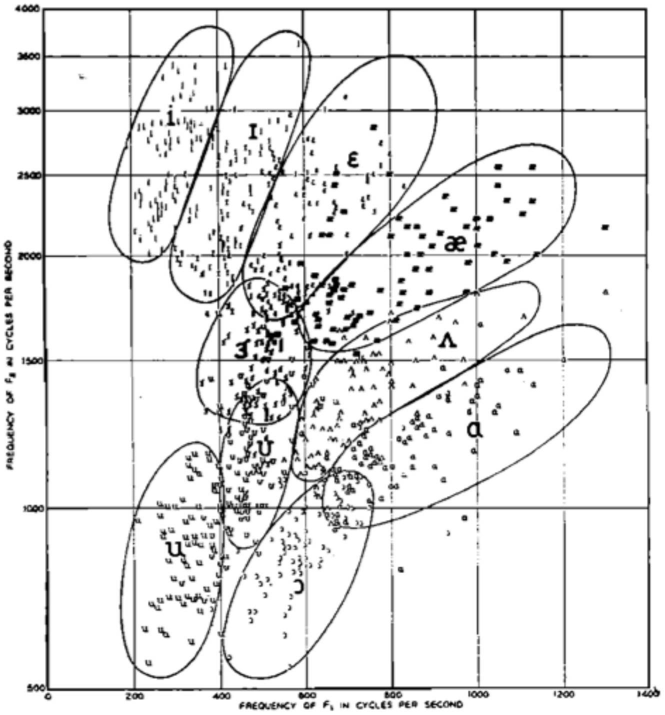


FIG. 8. Frequency of second formant *versus* frequency of first formant for ten vowels by 76 speakers.

Figure 2: A famous example of a well-studied phonetics-phonology map, the vowel formant plot (Peterson and Barney 1952).

> ellipses and not only the data points shown in Fig. 1 are mapped to a member (or multiple members) of the set of English vowel phonemes. Further, because the map is defined over a continuous space, it would never be possible to hear all the $\langle F1_{SS}, F2_{SS} \rangle$ points enclosed in the ellipses because there are infinitely many. Thus, *in learning a phonetic-phonological map defined over a* ***continuous*** *space, generalization occurs from a finite data sample to an infinite set*.

As Pierrehumbert (1990) discusses, phonetics-phonology representational maps have parallels to the "semantic" form-meaning representational maps in morphosyntax:

(3) Morphosyntax : {sequences of morphemes} → {sets of meanings}

There is ambiguity in form-meaning mappings in morphosyntax, too, especially when we abstract away from relevant context (e.g. pragmatic and prosodic context in morphosyntax; morphosyntactic context in phonetics-phonology); moreover, we add that generalization from a finite data sample to an infinite language occurs for both learning problems. The major structural difference between the phonetics-phonology and morphosyntax maps is that phonetics-phonology maps are defined in the real rather than the discrete domain.[1] It is because of this structural difference that the mathematical machinery for studying the two different maps diverges.[2]

Our current understanding of the phonetics-phonology map, after Pierrehumbert (2003a), is in fact an elaboration of (2): we augment each phonological category in the codomain with the probability that the sequence of phonetic parameter vectors belongs to it; we elaborate the map from one mapping $\langle F1_{SS}, F2_{SS}\rangle$ to sets of phonological categories, e.g.

(4) $$\langle F1_{SS} = 686, F2_{SS} = 1028\rangle \mapsto \{/\text{ɑ}, \text{ɔ}/\}$$

to a probability distribution of the categories over $\langle F1_{SS}, F2_{SS}\rangle$ vectors, e.g.:

(5) $$\langle F1_{SS} = 686, F2_{SS} = 1028\rangle \mapsto \{p(/\text{ɑ}/) = 0.45,\ p(/\text{ɔ}/) = 0.55\}$$

With our full model of the phonetics-phonology map as a **probabilistic distribution of phonological categories over a phonetic parameter space**, the key questions we need to answer to characterize the map are:

(Q1a) *What kinds of phonological categories are to be represented?*

(Q1b) *What is the phonetic parameter space for the phonological categories defined in (Q1a)?*

(Q1c) *What are properties of the distributions of the phonological categories of (Q1a) over the phonetic parameter space of (Q1b)?*

Phonological categories (Q1a) The choice of definition for the codomain of the phonetics-phonology map, the set of phonological categories, revolves around how contexualized the categories are. Peperkamp, Calvez, Nadal, and Dupoux (2006); Pierrehumbert (2003a,b) argue for the set to be a set of positional allophones, and for unification into phonemes using information from distributions of symbolic allophones or by using knowledge of the lexicon; Dillon, Dunbar, and Idsardi (Unpublished) argues for the set to be phonemes. Another option is to define the codomain over phonological features (Lin and Mielke 2008; Mielke 2008). Answering this question is not the focus of this paper, since we restrict attention here to tonal phonemes.

[1] There are also approaches to studying morphosyntax that model morphosyntax maps as being real-valued, cf. Widdows (2004): the co-occurrence of words in documents is used to determine similarity of word meanings, measured in real-valued vector spaces.

[2] It is possible to define a discrete phonetics-phonology map and thus study the phonological categorization problem using formal learning theory because we can represent the real-valued speech signal digitally to arbitrary precision in the limit, cf. Appendix A, (Jain, Osherson, Royer, and Sharma 1999). In fact, one may argue that the phonetics-phonology map is most correctly modeled over a discrete space because of precision limits in computing and biological systems (Blum 2004; Blum, Cucker, Shub, and Smale 1997). However, at the current stage of inquiry it is not clear how studying the phonetics-phonology learning problem using methods from formal learning theory gives us insight into how the learning occurs, and thus we do not pursue it here.

Phonetic parameter spaces (Q1b) It is answering (Q1b) that is the focus of this paper: to characterize the domain of the phonetics-phonology map by motivating which phonetic parameters are most significant for defining phonological categories; these are the dimensions that we want to define the distributions over. The set of phonetic parameters that may be extracted from the speech signal is obviously infinite in size and therefore must be constrained by some metric for computational tractability. For scientific purposes, too, we seek to limit the dimensionality of the phonetics-phonology map, i.e., the size of the parameter set, in order to have a succinct representation that is intelligible to the human scientist (Occam's razor). From the learner's perspective, a succinct learning target prevents overfitting to the input data and facilitates generalization to novel data (Duda, Hart, and Stork 2001:8-10), (MacKay 2003:343–349); from our scientific perspective, a succinct characterization of the representational map facilitates our ability to understand how the learning proceeds. In the best case, succinctness in the representational map results in no loss of information, i.e. without any smoothing out of the distributional modes corresponding to category structure in the phonetic space;[3] otherwise, the goal is succinctness with minimal loss of information.

We are in fact interested in characterizing three classes of phonetic parameter spaces to answer (Q0′), which is a question about lexical tone acquisition in general:

1. a universal parameter space $\mathscr{U}$ for all tone languages
2. the language-specific parameter space $\mathscr{L}$ for a given tone language
3. the speaker-specific parameter space $\mathscr{S}_{\mathscr{L}}$ for a given speaker of a given tone language.

By a parameter space, we mean the set of parameters over which the space is defined. By universal parameter space, we mean the smallest universal parameter space, the space which includes exactly and only the union of all language-specific parameter spaces.[4] To a first approximation, we assume:

(6) $$\forall \mathscr{L}, \forall \mathscr{S}_{\mathscr{L}},\ \mathscr{U} \supseteq \mathscr{L} \supseteq \mathscr{S}_{\mathscr{L}}.$$

This entails that the universal parameter space $\mathscr{U}$ can draw more distinctions than any tonal language-specific parameter space $\mathscr{L}$, which can, in turn, draw more distinctions than any speaker-specific parameter space for that language, $\mathscr{S}_{\mathscr{L}}$.

The assumption in (6) is motivated by the overarching idea based on empirical work on infant speech perception development over the past few decades that infants

[3] A simple example of succinctness without information loss is the expression of a finite language as a finite state automaton rather than as a list, since it takes fewer symbols to specify the finite state automaton than the list, and exactly and only the same sentences in the language are expressed (Meyer and Fischer 1971).

[4] The notion of a parameter space for all tone languages assumes that the class of tone languages is definable as a subset of all natural languages. Whether the restriction of languages to tone languages is available in acquisition is an open question, i.e. do children know they are learning a tone language, and if they do, under what conditions do they do this, and how do they do this? For the scope of this paper, we assume a restriction to a parameter space for tone languages for convenience.

begin as “citizens of the world” in having a universal ability to distinguish between sound categories and develop language-specific maps of the acoustic space through exposure to language input (Kuhl 2004). For instance, one of the first results of this kind was that English-learning infants showed behavioral responses consistent with the ability to discriminate between a velar stop (a sound in English) and a uvular stop (a sound not in English, but in Salish) at 6-8 months of age, but that by 10-12 months of age, they did not anymore (Werker and Tees 1984). Subsequent work confirmed and built on these results to flesh out a developmental timeline of perceptual reorganization of the acoustic space in which:

- Infants show a decline in their ability to discriminate nonnative vowel contrasts between 4-6 months (e.g. Polka and Werker 1994).
- Infants learning a non-tonal language show a decline in their ability to discriminate lexical tonal contrasts between 6 and 9 months (Mattock, Molnar, Polka, and Burnham 2008).
- Infants show a decline in their ability to discriminate nonnative consonantal contrasts between 6-8 and 10-12 months (e.g Werker and Tees 1984).
- Infants show improvement (facilitation) in their ability to discriminate native consonantal contrasts over the first years of life (Kuhl, Stevens, Hayashi, Deguchi, Kiritani, and Iverson 2006; Sundara, Polka, and Genesee 2006).
- Infants may be able to discriminate some native contrasts only after exposure to native language input[5] (Narayan, Werker, and Beddor 2010).
- A nonnative contrast that infants show a decline in discriminating can be learned by adult speakers of the same native language after significant exposure to the nonnative language (Tees and Werker 1984).

The cross-linguistic variability in the dimensions of acoustic spaces for phonological contrast and distributions of phonological categories over these spaces, as well as the change in the dimensions and distributions for infants due to language input show that *the phonetics-phonology map must be learned from language input*. It also motivates the need to study the phonetics-phonology map using cross-linguistic data to answer (Q0′).

The empirical evidence that: (i) language learners show decline rather than loss in sensitivity to particular phonetic dimensions, (ii) they can reactivate sensitivity

[5] Based on results like these, an alternative assumption to (6) is that

(7) $$\forall \mathscr{L},\ \mathscr{L} \supseteq \mathscr{U}.$$

based on the idea that sensitivity to some phonetic parameters may become activated only after exposure to language input. We do not take this alternative assumption because there is, to date, little supportive evidence for it. More importantly, a negative result for infant sensitivity to a speech sound contrast is conditional on a given experiment using a given task. A positive result is conditioned in the same way, as well, but shows that, at least under some conditions, infants show sensitivity to the contrast, while a negative result does not imply that infants are not sensitive to the contrast under any conditions.

with later language exposure and training, and (iii) listeners show the ability to use a wide variety of cues in degraded speech[6] suggests that the model of the development of language- and speaker-specific spaces of each language involves *parameter tuning/re-weighting* rather than *parameter selection*. Even in cases where sensitivity to some phonetic parameter may be vanishingly small, the model should assign it a vanishingly small weight rather than remove the parameter from the space.

Note that even for the purposes of studying the phonetic parameter space, we must represent data with a set of initial parameters: this initial set should be exactly $\mathscr{U}$, which we assume to be a superset of the dimensions of $\mathscr{L}$ for any natural tone language $\mathscr{L}$, cf. (6), and which is a subset of the set of all acoustic parameters we could extract from the speech signal. But these are not well-defined lower and upper bounds on $\mathscr{U}$; we cannot know what $\mathscr{U}$ is before studying what it should be! Thus, we make a guess and initialize the parameter set of $\mathscr{U}$ based on cross-linguistic work on tonal production, perception, and automatic tonal recognition.

The distribution (Q1c) We assume that the distribution of phonological categories over the phonetic space is continuous. Since the details of the distribution depends strongly on how the phonological categories and the phonetic space is defined, we let our study of those determine characteristics of the distribution. These characteristics then inform how we constrain the type of distributions available in the hypothesis space for the learner in modeling the actual learning of the representational map.

1.2 *Methodological Abstractions*

In characterizing the phonetic parameter space (Q1b) for lexical tonal categories in this paper, we make three main methodological abstractions: (i) to sharpen the probabilistic distributions of phonological categories into partitions over the phonetic space, (ii) to use category separability as a metric for constraining the phonetic parameter space, (iii) to limit the context available to extract the phonetic parameters from, and (iv) to introduce linguistic structure into the unanalyzed speech signal. Characterizing the phonetic parameter space with these methodological abstractions in place still allows us to bear on questions (Q1a)–(Q1c).

Partitions over the phonetic space and category separability While the reality is that the phonetics-phonology map is a probabilistic distribution of phonological categories over the phonetic space, in characterizing the phonetic parameter space, we make the methodological abstraction that the map is a *partition* of phonological categories over the space: every point in the space maps to exactly and only one phonological category.

The reason for the abstraction is that most well-understood computational algorithms for classification give "hard" classifications, i.e. produce a partition of the space, rather than a probabilistic distribution over it (Wahba 2002). Moreover, while it is possible to elicit probabilistic confidence ratings in human perception experi-

[6] see Assmann and Summerfield (2004) for a general review of perception of degraded speech.

ments, e.g. using magnitude estimation (Bard, Robertson, and Sorace 1996), we use forced choice tasks in our perception experiments to match the hard classification of the computational algorithms.

Along with the methodological abstraction of modeling the phonetics-phonology map with partitions, we use the general metric of *category separability* to determine how relevant/informative phonetic parameters are for defining the tonal categories: more informative phonetic parameters define a space in which the tonal categories are better separated. As discussed by Nearey (1989), this category separability metric is *data analytic* because it is based on production data only, while ultimately, *perceptual* separability from listening experiments is what is directly relevant for the representational map. However, data analytic category separability certainly bears on perceptual separability.

Limiting context for phonetic parametrization We have already proposed a map (2) restricting the domain to phonetic parameters. We reiterate here that we are abstracting away from non-phonetic context, e.g. morphosyntactic information (the language model in automatic speech recognition), to constrain the research problem; Jansen (2008) calls this the "pure speech" setting. Moreover, we restrict the *temporal domain* for phonetic parameter extraction. The strongest such restriction is to restrict the extraction of phonetic parameters to only the unit to the classified, e.g. only from the syllable of the tone to be classified. In this paper, we start from this restriction, but we will ultimately allow parameter extraction from the preceding and following syllables as well. For fluent speech recognition, there is strong evidence that humans extract parameters from temporal domains wider than the unit to be classified, e.g. Ladefoged and Broadbent (1957); Wong and Diehl (2003).

Introducing linguistic structure in the speech signal While the original research question (Q0′) assumes extraction of parameters from the unanalyzed signal, for this paper, we extract parameters from speech segmented for syllabic structure for convenience. This is like having an oracle tell the classifier where syllable boundaries or onset/rime boundaries are. In future work, we can remove this extra information by implementing a sonority detector to find syllables, as in Jansen (2008).

1.3 *The Model System for the Acquisition of Lexical Tones*

With the larger research questions and the methodological abstractions set up, we turn to the model system under study.

The gross characterization of our model system is this:

- **Data**: monotones extracted from sentence-medial position in connected speech over cross-linguistic tonal language sample
- **Phonetic parameter space**: acoustic parameter space, extracted from the speech signal
- **Phonological categories**: lexical tonal phonemes (tonemes)

Like any other system studied in phonological category acquisition, the one we study here is a model system, and we study it with the same scientific motivation that a biologist studies a simple model organism like baker's yeast (the eukaryote with the smallest number of genes) to illuminate gene regulation in more complex systems such as humans (Fields and Johnston 2005). Clearly the model system can only capture certain aspects of the process of phonological category acquisition, highlighting some while muting others. In this section, we describe how we instantiate the model system for lexical tone acquisition to answer (Q0′).

Our research questions, as laid out in §1, dictate the following requirements for setting up a model system for studying lexical tone acquisition:

- A representative cross-linguistic sample to address the language-specific development of speech categorization
- A language sample relevant for modeling language input to infants
- Some controlled source(s) of variability to enable modeling the challenge of categorization in the face of variability

Cross-linguistic tone language sample We chose a sample of tonal languages to include: (i) register/level tone languages, with only level tones (Bole, Igbo), and (ii) contour tone languages with contour tones and level tones (Mandarin, Cantonese, Hmong)[7]. We summarize the diversity of the cross-linguistic tonal language sample below in Table 2, using International Phonetic Alphabet notation for the tonal inventory, and give recording details of the data currently available below in Table 3.

Language	Area	Tonal inventory	Phonation
Bole	Nigeria	˥, ˩ (H,L)	
Igbo	Nigeria	˥, ˧, ˩ (H, !H, L)	
Mandarin	Beijing, Taiwan	˥, ˧˥, ˨˩˦, ˥˩	creaky ˨˩˦, ˥˩
Cantonese	Hong Kong	˥, ˦, ˧, ˨˩, ˨˧, ˧˥	creaky ˨˩
Hmong	Laos/Thailand	˥, ˧, ˨, ˥˨, ˦˨, ˨˩, ˨˦	breathy ˦˨, creaky ˨˩

Table 2: Cross-linguistic sample of tonal languages recorded to provide language input

Language	Dialect	Recording location	Speakers
Bole	Fika	Potiskum, Nigeria	3M/2F
Igbo	Anambra	Los Angeles, CA	1M/2F
Mandarin	Beijing	Beijing, China	6M/6F
Mandarin	Taiwan	Los Angeles, CA	6M/6F
Cantonese	Hong Kong/Macau	Los Angeles, CA	6M/6F
Hmong	White	Fresno, CA	6M/5F

Table 3: Details for recordings of language sample

[7] The language sample was also chosen to exhibit a variety of tone-voice quality interactions. While beyond the scope of this paper, our cross-linguistic data and perception experiments suggest that the parameterization of the speech signal for tonal representation must include voice quality parameters, e.g. related to phonation, beyond simple f0-based parameters, cf. Lam and Yu (2010); Yu (2010).

Language input to infants and sources of variability Because infants exhibit perceptual knowledge before articulatory knowledge of speech sound categorization (Kuhl 2004), we restricted parameters to *acoustic parameters* and abstracted away from articulatory parameters.

Other work on learning tonal categories has emphasized that the majority of the input to the infant consists of multiple words so that contexual variation due to tonal coarticulation from neighboring tones is a regular part of the input the learner receives (Gauthier et al. 2007; Shi in press). Specifically, Gauthier et al. (2007); Shi (in press) claim that about 90% of parental speech to infants is multi-word utterances. Moreover, the majority of language data an infant hears is not speech directed to the infant, but, for instance, adult-to-adult speech. An estimate from van de Weijer (1998, 2002) is that only about 14% of the input is direct speech to the infant.

Because of the large amount of input that infants hear that is adult directed speech and multi-word utterances, Gauthier et al. (2007) modeled learning tone categories based on speech from adults rather than infant-directed speech, (and in general, research building tone recognizers is modeled on adult speech). This is of course a working hypothesis; surely the presence of infant directed speech and isolated words in the input could affect the character of the learning problem.[8] We follow this choice, taking our input to the learner to be adult connected speech. We capture the role of contextual tonal variation in creating variability in the input by collecting the full permutation set of bitones in connected speech for each language in the sample, and we capture interspeaker variation by recording multiple speakers of both genders.

This concludes our section on preliminaries, which we have deliberately kept broad in scope to illustrate our model system of lexical tone in context of the study of phonological (and language) acquisition in general. We now turn to describing our exploration of the two issues regarding f0 parametrization discussed in the introduction: coarse temporal resolution in parameterization and static and dynamic parametrizations of f0.

2 The Parametrization of f0 in Representational Maps for Lexical Tone

Gauthier et al. (2007), the only preceding computational modeling study of learning a tonal system (the four basic Mandarin tones), suggests that representing examples to the learner as densely sampled f0 velocity contours results in more robust tonal categories than representing examples as densely sampled f0 contours.

[8] For instance, note that the rationale for the ecological validity of adult connected speech given above assumes equal weighting in infant attention to all input regardless of whether it is directed to the infant. In fact, studies show biases for infant directed speech over adult speech and biases for the infant for their mother's voice and the importance of placing language input within social interaction (Kuhl, Tsao, and Liu 2003). Thus, it is not unreasonable to hypothesize that despite the relatively small amount of infant directed speech in the ambient input, it may be a rich source of information for infants about learning tone patterns. In fact, work has found correlation between the amount of exaggeration in infant directed speech in terms of the expansion of the vowel and tonal spaces in predicting an infant's ability to discriminate native consonant contrasts (Liu, Kuhl, and Tsao 2003; Xu and Burnham submitted).

Moreover, the study suggests that *parametrization of the speech signal as densely sampled f0 velocity contours (f0′) alone is sufficient for learning Mandarin tones*. The intuition for why f0 velocity might be more relevant than f0, and furthermore, sufficient alone for tonal classification, is that the derivative of a constant function is zero: f0 velocity provides a way of speaker normalization, of removing constant shifts due to different pitch ranges.

We generalize this hypothesis as an initialization for $\mathscr{U}$: $\mathscr{U}_G$ is a d-dimensional parameter space defined over d densely, uniformly spaced f0 velocity (f0′) samples from the syllable; each of the d samples contributes a dimension to the space, and the sampling rate is defined over time normalized by the syllable duration, t_{syll}, i.e., a sample taken at timepoint $t_{syll} = i$ is taken at $i/(d-1)$ of the way through the syllable:

(8) $$\mathscr{U}_G = \{\mathrm{f0}'(t_{syll} = i) \mid 0 \leq i \leq d-1, d \text{“large”}\}$$

$\mathscr{U}_G$ assumes dense temporal sampling resolution and a parameterization including only a dynamic f0-based parameter.

We hypothesize, in contrast, that:

(H1) *Coarse temporal sampling resolution of the parameterized speech signal is sufficient for good tonal category separability.*

(H2) *The parametrization of the speech signal as f0 velocity contours is not sufficient for good tonal category separability cross-linguistically.*

2.1 *Coarse Temporal Resolution (H1)*

Increasing temporal resolution means increasing the dimensionality of the parameter space: each additional sample adds a dimension. Thus, coarse temporal resolution is necessary for a succinct tonal representation, which is desirable for generalization in learning a phonetics-phonology and for scientific understanding, cf. §1.1.

Linguistic models for the representation of tone implicitly advocate coarse temporal resolution. Chao (1930)'s tone letters used in the International Phonetic Alphabet for representing tones, e.g. ˨˩˦, suggest that three samples (and more specifically, three particular samples) over the tone are sufficient, as described in Chao (1968)'s model of Chinese tone systems in his grammar of Chinese:

> If we divide the range of a speaker's voice into four equal intervals, marked by five points, 1 low, 2 half-low, 3 middle, 4 half-high, and 5 high, then practically any tone occurring in any of the Chinese dialects can be represented unambiguously by noting the beginning and ending points, and, in the case of a circumflex tone, also the turning point; in other words, the exact shape of the time-pitch curve,

> so far as I have observed, has never been a necessary distinctive feature, given the starting and ending points, or the turning point, if any, on the five-point scale. (Chao 1968:25)

The *modus operandi* in speech recognition, though, is to use a constant frame rate, sampling features every 10ms over 30ms windows (Young, Evermann, Gales, Hain, Kershaw, Liu, Moore, Odell, Ollason, Povey, Valtchev, and Woodland 2009), and Gauthier et al. (2007)'s sampling rate (30 samples/syllable) is close to this.

However, a survey of sampling characteristics in the automatic tonal recognition literature suggests that coarse sampling of f0 parameters can yield good performance, as summarized in Table 4 below. In the table, we also indicate the *clock* used for each study, by which we mean which temporal unit was used to define the (uniform) sampling rate. For the studies where we describe the sampling in terms of "slices", this means that features were extracted as averages over the slices, i.e. smoothed.

Study	Language	Clock	Sampling resolution
Zhang and Hirose (2004)	Mandarin	Absolute time	Fine, 10ms frame shift
Gauthier et al. (2007)	Mandarin	Normalized to syllable	Fine, 30 samples/syll
Ọdẹ́lọbí (2008)	Yoruba	Normalized to syllable	Medium, 9 slices/syll
Wang and Levow (2008)	Mandarin	Normalized to tone nucleus	Coarse, 5 samples/nucleus
Qian, Lee, and Soong (2007)	Cantonese	Normalized to rime	Coarse, 3 slices/final
Zhou, Zhang, Lee, and Xu (2008)	Mandarin	Normalized to nucleus	Coarse, 3 slices/nucleus

Table 4: Sampling characteristics of a selection of tone recognition studies

The predominance of coarse sample resolution and linguistically-tied clocks in recent tonal modelling is very striking, compared to the predominance of high frame rate and absolute time in sampling in general speech recognition. Note that no study sampled fewer than 3 times per tonal domain.[9] One automatic tonal recognition study of Mandarin even found that coarse sampling, with 4 samples/tone, outperformed dense sampling with 1 sample/10 ms (Tian, Zhou, Chu, and Chang 2004).

In summary, long-standing linguistic intuition and evidence from recent large-scale automatic tonal recognition studies converge to suggest that coarse sampling is sufficient in parametrization of tonal spaces. In our research, we confirm this with experimental and computational modeling work: (i) a human tonal perception experiment studying the effect of sampling resolution on Cantonese tonal perception and (ii) computational studies of the effect of sampling resolution on category separability over our cross-linguistic tonal sample.

[9] For Mandarin at least, the reason why is hinted at already in Chao (1968): "practically any tone occurring in any of the Chinese dialects can be represented unambiguously by noting the beginning and ending points, and, in the case of a circumflex tone, also the turning point." Two f0 feature samples is not sufficient to distinguish Tone 2 (rise) and Tone 3 (fall-rise) in isolation. Zhou et al. (2008) empirically studied this in their multilayer perceptron Mandarin tone recognizer: in varying the number of inputs to the neural network, they found that percent correct saturated after the number of inputs was increased from 2 to 3, and that the improvement was due to improvements in classification from Tone 3.

2.2 *Insufficiency of f0 Velocity Contours (H2)*

The second hypothesis is that f0 velocity contours, regardless of sampling resolution, are insufficient for good separability of tonal categories. The obvious counterexamples to an initialization $\mathcal{U}_G$ in (8) are a level/register tone language and a tone language with level and contour tones. The Mandarin tonal inventory that is the target of learning in Gauthier et al. (2007) is unusual in having no level tone contrasts.

We note that the level tone counterexample is not trivial, i.e. it is not enough to reject $\mathcal{U}_G$ with a thought experiment. Level tone sequences are not a series of step functions, but may in fact be realized as if they are contour tone sequences due to contextual tonal variation, cf. Figure 3 and Maddieson (1977:337).

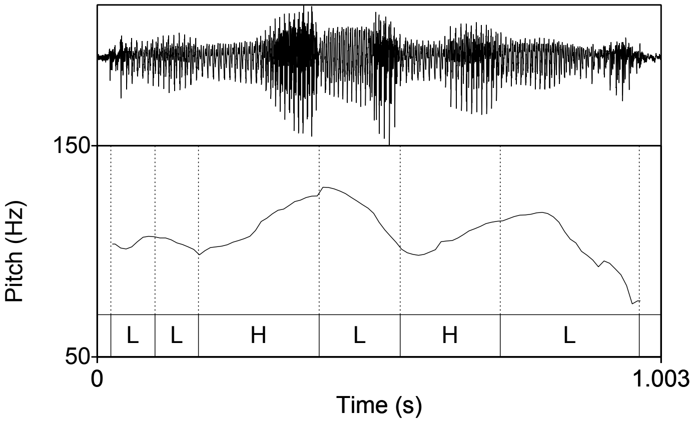


Figure 3: A sequence of tones in Bole, a tone language with H and L tones. Sequences of level tones in a level tone language are not necessarily sequences of step functions. Rather, they can show rises and falls due to tonal coarticulation. The sentence is *ànìn némà méngò*, 'The owners of prosperity came back.'

If $\mathcal{U}_G$ was the structure of the universal parameter space for tones, we might expect many tonal systems to consist of purely dynamic contrasts. In fact, a striking typological pattern in tonal inventories is that two-tone systems of this kind are not known to exist, as noted as early as the 1960s:

> The simplest language of [a pure contour tone system] would have two tonemes, one a glide upwards and one a glide downwards, with the level of the end points of complete irrelevance to the system. Here the contrast would be that of a rising contour opposed to a falling contour. No system this simple has come to my attention. (Pike 1964:9)

Instead, the dominant tonal system is an inventory of two level tones, H and L, like Bole in our language sample; in the statistical sample of tone languages in Maddieson (1978), about half of the languages are of this type.

To test the relevance of f0 velocity contours, we compare tonal category separability in our modeling using: (i) only f0 velocity (ii) only absolute f0, and (iii) both f0 velocity and absolute f0. Given our goal of modeling human cognition, it would be useful to study how human tonal perception proceeds when only f0 velocity cues are present. However, factoring out all cues in the speech signal, including f0 height, except f0 velocity is not possible, although attempts of this kind have been made by psychophysicists (Dooley and Moore 1988; Divenyi 2004). Thus, we confine our studies bearing on (H2) to computational modeling studies.

3 Experimental and Computational Studies Bearing on the Parameterization of f0 in Tonal Spaces

In this section, we briefly summarize results from our own experimental and computational work bearing on the hypotheses H1 and H2. First, we discuss experimental results showing that human listeners can maintain tonal identification accuracy with stimuli degraded to be coarsely sampled (§3.1). Then we discuss exploratory computational studies of the parameterization of tonal spaces using coarse and dense sampling of absolute f0 and f0 velocity (§3.2).

3.1 Coarse Temporal Sampling and Human Tonal Perception

In a Cantonese tonal perception experiment in which we manipulated the sampling resolution in the stimuli presented to the listener, we showed that tonal identification accuracy under coarse temporal sampling down to 3 samples/syllable can be as high as accuracy with the intact signal.

Cantonese tritones ⟨*wai*˧, {*wai*˥, ˧˥, ˦, ˨˩, ˨˧, ˧} , *mat*˧ ⟩ extracted from connected speech by multiple speakers (3 M, 2 F) were presented to 39 native Cantonese listeners in sound-attenuated booths at City University of Hong Kong and UCLA.[10] The listeners were asked to identify the second tone in the tritone by a key press of the corresponding orthographic label. Sampling resolution varied from the intact signal, to 7, 5, 3, and 2 30.4-ms uniformly spaced samples (time-slices) per syllable. The stimuli were blocked by sampling resolution, and block order was pseudorandomized to be roughly uniformly distributed over sampling resolution.

The sampling resolution manipulation involved intermittently replacing the speech with noise 10dB higher than the signal amplitude, as in multiple phonemic restoration (Bashford, Riener, and Warren 1992; Miller and Licklider 1950), cf. Figure 4, using Matlab and Praat (Boersma and Weenink 2010).

A repeated measures ANOVA with SAMPLING RESOLUTION as a fixed effect and SUBJECT as a random effect showed a main effect for SAMPLING RESOLUTION: $F(4, 152) = 28.6, p < 2.2 \times 10^{-16}$. Bonferroni corrected pairwise comparisons with the familywise Type I error rate at 0.05 showed significant differences between the 2-sample condition and all other conditions, and between the 3-sample condition and the 7-sample and intact conditions. Thus, on average, listeners were able to maintain tonal identification accuracy down to 5 samples/syllable, and also, to some degree,

[10] Tritones rather than monotones or bitones were used to preclude a floor effect washing out any differences between sampling resolution conditions.

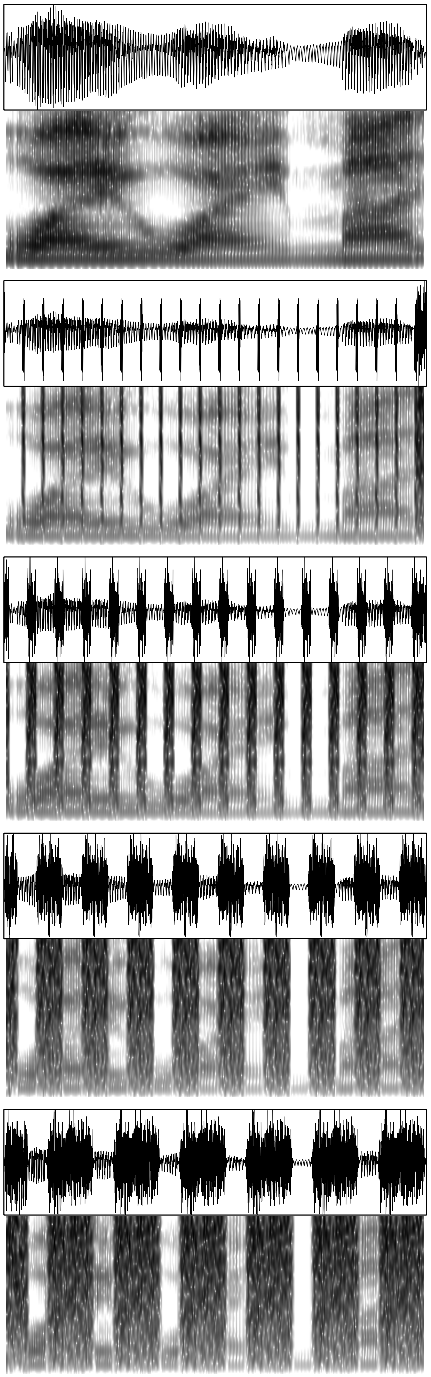


Figure 4: Waveforms and spectrograms of sample stimuli for sampling resolution from intact, to 7, 5, 3, and 2 samples/syllable over Cantonese tritones

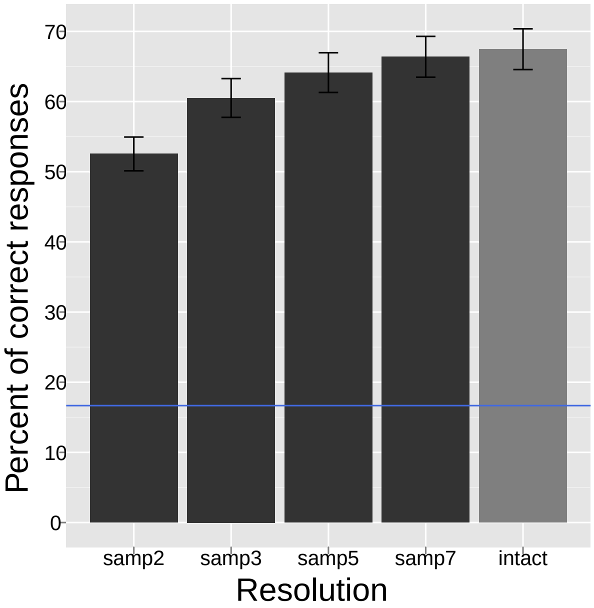


Figure 5: Comparison of tonal identification accuracy for different sampling resolutions. Tonal identification accuracy was maintained from the intact signal down to 3 samples/syllable. For all sampling resolutions, performance was also well-above chance (the blue line shows identification accuracy for at-chance performance (1/6)), and the error bars show ±1SE.

down to 3 samples/syllable, but not down to 2 samples/syllable, cf. Table 5 and Figure 5.

Resolution	Percent correct (SE)
samp2	52.54 (2.41)
samp3	60.51 (2.76)
samp5	64.13 (2.83)
samp7	66.38 (2.91)
intact	67.46 (2.9)

Table 5: Tonal identification accuracy for different sampling resolutions averaged over the listeners.

The Cantonese tonal perception results therefore support the hypothesis that coarse temporal resolution may be sufficient for good tonal category separability. However, the experimental results do not inform us as to what cues the listeners are using in those few samples to identify the tones with reasonably high accuracy.

3.2 *Computational Modeling and the Parametrization of f0*

In this section, we briefly summarize results from initial computational modeling bearing on hypotheses H1 and H2, regarding category separability under dense and

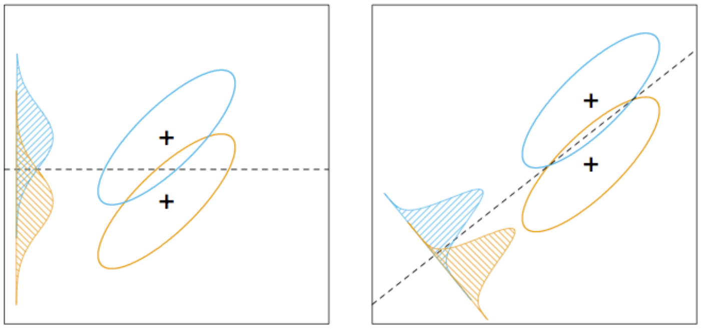


Figure 6: A geometric characterization of linear discriminant analysis (LDA) for a two-class problem, from Hastie et al. (2009:116). The objective is to maximize the ratio of the between-class variance to the within-class variance. Thus, though the projection in the left panel along the direction of the line connecting the centroids maximizes the between-class variance, the within-class variance is high and there is large overlap in the two classes. The projection on the right minimizes the ratio of between- to within-class variance and is the projection chosen in LDA.

coarse sampling, and the insufficiency of f0 velocity contours for good category separability. As an initial measure of category separability, we choose linear discriminant analysis to aid in exploratory visualization of the multidimensional parameter space.

3.2.1 Linear Discriminant Analysis (LDA)

Linear discriminant analysis is both a dimensionality reduction technique and a classification algorithm (Hastie, Tibshirani, and Friedman 2009:§4.3). As a dimensionality reduction technique, it chooses a projection of the data into a smaller-dimensional space such that the projection maximizes category (class) separability, where the class separability is measured as the ratio of the between-class variance (the variance of the projected class means) to the within-class variance in the projected data (the pooled variance about these means), cf. Figure 6.

As a classification algorithm, it defines a partition of the space by estimating linear decision boundaries and classifies an observation into the class with the nearest centroid, measured by Mahalanobis distance (a distance metric that is covariance-adjusted). Under strong (and typically false) assumptions about the distribution of the data, namely, that the distribution of data within each class is multivariate Gaussian with a common covariance matrix, linear discriminant analysis is equivalent to a Bayesian classifier (Hastie et al. 2009:439).

We are primarily interested in using linear discriminant analysis for the purposes of exploratory visualization of data in low dimensions because our data, unsurprisingly, fail to satisfy the assumption of multivariate normality with common covariance matrices, and because we are interested in trying methods that allow more complex decision boundaries than linear decision boundaries.

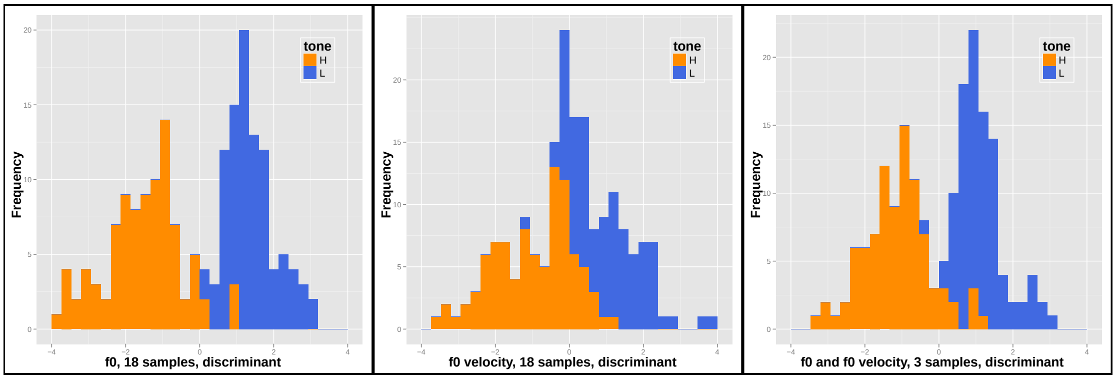
3.2.2 *Category Separability Under Coarse and Dense Sampling of f0-based Parameters*

Our initial modeling extracts parameters from only the tone to be classified, without any contextual information from neighboring tones, and we examine category separability for tonal spaces from single speakers.

By using LDA as implemented in R (R Development Core Team 2010) by Venables and Ripley (2002) to visualize our data, we compared category separability under coarse and dense sampling of: (i) absolute f0, (ii) f0 velocity, and (iii) both absolute f0 and f0 velocity. We calculated f0 averaged over coarse and finely divided uniform subsections (time slices) of the syllable using VoiceSauce (Shue, Keating, and Vicenik 2009) and calculated f0 velocity by taking differences between these averaged f0 values in R. To avoid linearly dependent parameters in parameter sets including both absolute f0 and f0 velocity, we calculated absolute f0 and f0 velocity separately with differing coarseness of subsection division, since the number of f0 velocity samples calculated from some number of absolute f0 samples is necessarily less than the number of absolute f0 samples.

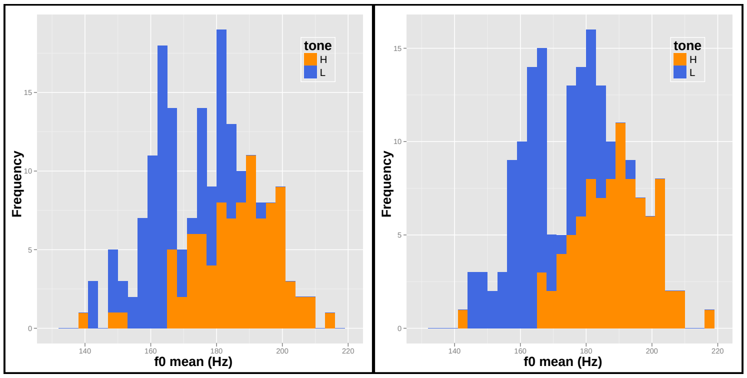
In general, our exploratory results suggest that for phonetic parameterization without any contextual parameters in single speaker spaces, coarse sampling (3 samples each) of absolute f0 and f0 velocity is sufficient for good category separability, and critically, category separability for coarse sampling of absolute f0 and f0 velocity is comparable or better than dense sampling (18 samples) of f0 velocity. Below, we exemplify our results with two examples.

In Figure 7, we show the phonetic parameter space for a Bole female speaker for densely sampled f0 velocity (18 samples), densely sampled absolute f0 (18 samples), and coarsely sampled absolute f0 and f0 velocity (3 samples each) after LDA dimensionality reduction. It is clear that category separability for densely sampled f0 velocity (Figure 7b) is poorest, while category separabilty for coarsely sampled absolute f0/f0 velocity (Figure 7c) is comparable to that for densely sampled absolute f0 (Figure 7a).



(a) Densely sampled absolute f0 (b) Densely sampled f0 velocity (c) Coarsely sampled absolute f0 and f0 velocity

Figure 7: The separability of Bole tones (H, L) for a single female speaker with coarse f0 and f0 velocity is similar to that with dense f0 sampling and better than with dense f0 velocity sampling.



(a) Mean absolute f0 (b) 1 absolute f0 sample

Figure 8: Category separability of Bole tones for a single female speaker with mean absolute f0 only (8a), or one sample of f0 at the midpoint from a 9-sample subdivison of the syllable (8b).

It is important to note that f0 velocity provides any evidence of category separability at all: this shows that it may be a relevant parameter for even the simplest of level tone systems, and more generally, that dynamic properties of f0 are relevant for level tones, contrary to characterizations of level tones that suggest one f0 height sample is enough to specify a level tone, while contour tones require multiple samples:

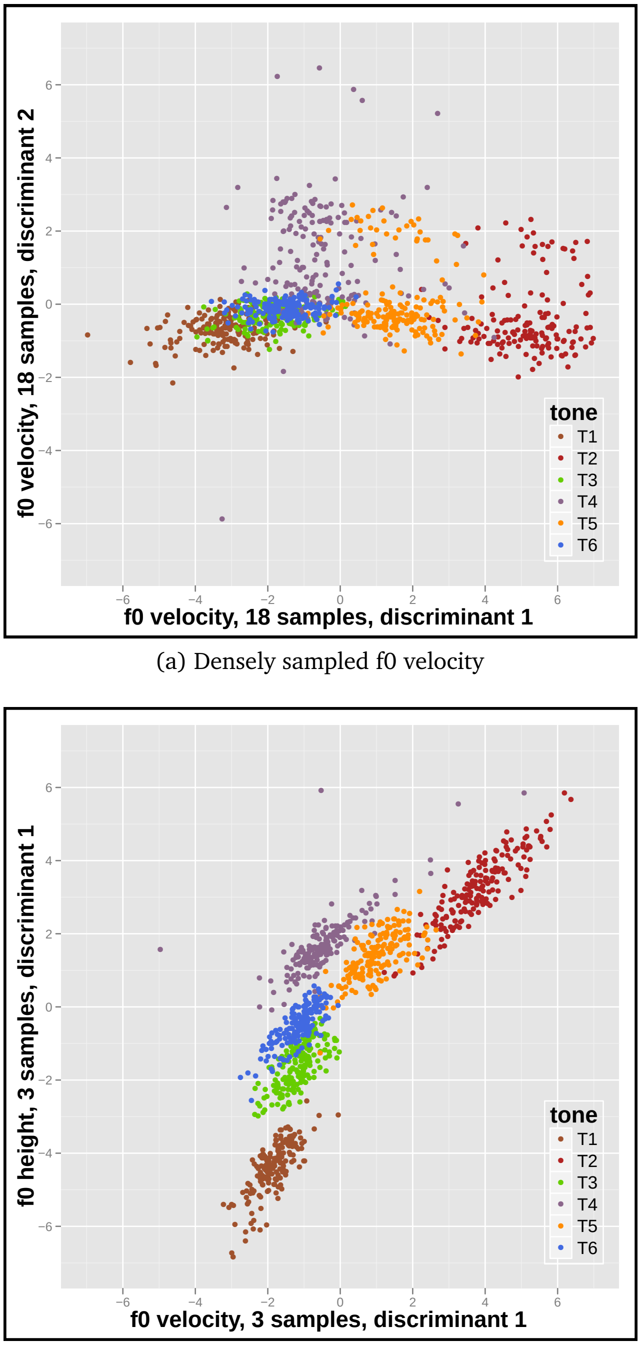
> If an adequate synthesis of a tone can be made by specifying a single level, it may be considered a level tone. But a tone represented by a pitch glide which cannot be generated by rule from the environment (i.e. not by a default) requires specification of several points. (Maddieson 1977:337)

In fact, category separation for the same speaker, using either only mean absolute f0 or one sample of absolute f0 from the midpoint is poor, cf. Figure 8, and the bimodal distribution of the L tone suggests latent category structure not captured by the 1-dimensional parametrization, at least without a relational parameterization of f0.

Results in a tonal system with many contours and levels, such as Cantonese, are even stronger: with densely sampled f0 velocity contours (18 samples), the three level tones show gross overlap, cf. Figure 9a. However, with coarse sampling of absolute f0 and f0 velocity (3 samples each), cf. Figure 9b, a near partition of the tonal categories, levels and contours, is obtained.

These results support Hypothesis H2: f0 velocity, regardless of density of sampling, is not sufficient for good category separability across all languages. However, coarse sampling of relevant parameters—both absolute f0 and f0 velocity—results in a near-partition of the phonetic parameter space.

While highly preliminary, our modeling results using linear discriminant analysis suggesting that coarse sampling of relevant features suffice for good category separability converge with results from automatic tonal recognition (Tian et al. 2004) and our own experimental work (§3.1), and extends those results to a larger range of languages.



(a) Densely sampled f0 velocity

(b) Coarsely sampled absolute f0 and f0 velocity

Figure 9: Category separability for the parameter space for a single male speaker of Cantonese. Cantonese level tones (Tones 1, 3, 6) cannot be separated with f0 velocity information alone, but coarsely sampled absolute f0 and f0 velocity parameterization results in a near-partition of the phonetic space.

We are currently implementing studies on the parameterization of the speech signal using multiclass support vector machines (Crammer and Singer 2001) and maximum entropy methods (Pietra, Pietra, and Lafferty 1997). Support vector machines are well-understood algorithms that, like LDA, calculate optimal separating hyperplanes (linear decision boundaries, i.e. lines in 2-D spaces) over a parameter space for separating classes. Moreover, they effectively allow the calculation of more complex, nonlinear decision boundaries in the parameter space by efficiently calculating optimal separating hyperplanes in higher-dimensional parameter spaces. Maximum entropy methods have the desirable property that the search space for optimizing the parameter set is convex. This means that we can avoid getting trapped in local minima in the search space, i.e. on a hypothesis for the characterization of the parameter set in which all small movements in the search space result in a less optimal hypothesis.

Conclusion

In conclusion, our work thus far suggests that coarse temporal sampling resolution of the parametrized speech signal is sufficient for good tonal category separability, given that relevant, informative parameters are sampled. This is supported by exploratory modeling using linear discriminant analysis across the level and contour languages of our sample and by the maintenance of tonal identification accuracy under stimuli degraded to be coarsely sampled in a Cantonese tonal perception experiment.

Our modeling work also suggests that the parametrization of the speech signal as f0 velocity contours alone is not sufficient for good tonal category separability cross-linguistically. The strongest evidence for this is the immense overlap of level tone categories in mixed level tone/contour tone systems in a phonetic parameter space defined only over f0 velocity, even for a single speaker. Category separability in supervised learning gives an upper bound for category separabilty in unsupervised learning; the failure of Cantonese level tones to be separated in a pure f0 velocity space with LDA therefore implies that linear clustering methods cannot succeed, either. Thus, Gauthier et al. (2007)'s suggestion that f0 velocity is sufficient for learning tones cannot be generalized cross-linguistically, although we point out that f0 velocity is a relevant parameter for level tone systems, a reflection of the fact that level tones can be realized as contours due to tonal coarticulation.

Our future work will continue to home in on a definition of the target of learning in the acquisition of tones from the speech signal. With a definition of tones that is well-motivated by what we know about human tonal production and perception, we can model how lexical tones could be learned from the speech signal in a way consistent with what we know about human cognition.

Acknowledgements

We would like to acknowledge Edward Stabler, and Meaghan Fowlie, Thomas Graf, Bruce Hayes, Patricia Keating, Russell Schuh, Megha Sundara, Colin Wilson, Kie Zuraw, and audiences at the UCLA Phonetics Seminar and NELS41 for illuminating discussions and feedback, Hiu Wai Lam for testing the perception experiment subjects, Eric Zee for letting us run experiments in his laboratory at City University of Hong Kong, all those who helped with linguistic consultation and recordings: Russell Schuh and Alhaji Maina Gimba (Bole), Echezona Okoli and Moses Chikwe (Igbo), Mandarin (Jianjing Kuang), Hiu Wai Lam, Shing Yin Li, Cindy Chan, Cedric Loke, and Vincie Ho (Cantonese), Chou Khang and Phong Yang (White Hmong), and all the dedicated undergraduates that helped with phonetic data processing: Samantha Chan, Chris Fung, Hiu Wai Lam, Wai-Ting Lam, Antonio Sou, Kristen Toda, Grace Tsai, and Joanna Wang. This work was supported by a NSF graduate fellowship, a Ladefoged Scholarship from the UCLA Department of Lingustics, and NSF grant IIS-1018863 to Abeer Alwan, Patricia Keating, and Jody Kreiman. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

A The phonetics-phonology map learning problem in formal learning theory

As stated in Footnote 2 on page 5, it is possible to define a discrete phonetics-phonology map and thus study the phonological categorization problem using formal learning theory because we can represent the real-valued speech signal digitally to arbitrary precision in the limit.

We assume the phonetics-phonology map is a *partition* of a finite set of phonological categories, *Cat*, over the phonetic parameter space:

Phonetics-Phonology:

(9) {sequences of phonetic parameter vectors} → {phonological categories}

Choose any phonetic parameter vector $\vec{v} \in \mathbb{R}^n$, for finite n, e.g. $\langle F1, F2\rangle \in \mathbb{R}^2$. Digitize $\vec{v}$ with an n-bit quantization, where $n \in \mathbb{Z}$, and sample at some sampling rate s.[11,12] Then at each timepoint t of sampling, each entry of $digitized(\vec{v})(t)$ is in PnF, where PnF is the finite set of 2^n different symbols from the n-bit quantization.

Consider the language L which is a set of pairs: $\langle p, c\rangle$ where $p \in PnF$, $c \in Cat$, and assume L is in the class of r.e. languages. We would like to show that the class of such languages, $\mathscr{L}_{PnPh}$, is learnable by constructing a learner $\phi : (\mathbb{Z}^n \times Cat) \mapsto \mathscr{G}$ to map from the class of languages to a class of grammars.

By assumption that the phonetics-phonology map is a partition, (i) each p is paired with a unique c, i.e. L is a function (single-value language), and (ii) $\forall p$, p is mapped to some c, so this function is total. Thus, $L \in \mathscr{L}_{svt}$, the class of total

[11] Note that discrete phonetic parameterization (beyond digital speech processing) is not unusual, e.g. fundamental frequency is often parametrized as 5-valued (Chao 1930).

[12] This representation of the speech signal is based on Pierrehumbert (1990:379).

single-value languages and therefore the class of languages of phonetic parameter vector-phonological category maps is identifiable from positive data (Jain et al. 1999).

The learner for $\mathscr{L}_{svt}$ in Jain et al. (1999) is not computable as it relies on an enumeration of the grammars of all r.e. languages. Thus, even though recognizing that $\mathscr{L}_{PnPh} \subseteq \mathscr{L}_{svt}$ is sufficient for proving it is learnable (in the Gold sense), we may argue that this learnability result doesn't reveal the structure of the learning problem for phonetics-phonology maps.

However, we might also argue that the particular class of languages relevant for the phonetics-phonology map is a *proper subset* of the class of total single-value languages: $\mathscr{L}_{PnPh} \subset \mathscr{L}_{svt}$. In particular, we can assume a fixed finite phonetic parameter set for $\vec{v}$, a fixed n-bit depth for quantization, a fixed sampling rate s, and a finite, fixed set of phonological categories *Cat*. With these fixed bounds, $\mathscr{L}_{PnPh}$ is a finite subset of the finite languages. Since $\mathscr{L}_{PnPh}$ has finite cardinality, the VC dimension of this class is also finite and thus $\mathscr{L}_{PnPh}$ is PAC-learnable: the class of languages for phonetics-phonology maps is both computable and tractable.

Even with this result for computability and feasibility, we don't pursue a finite model for the phonetics-phonology learning problem. Although there may be grounds to model the speech signal with a discrete representation based on finiteness in the number of distinctions that human sensory systems can draw, that finiteness is vast: the cardinality of $\mathscr{L}_{PnPh}$, even if finite, is of a vastness of the order of magnitude so that idealization of phonetic parameterization as being real-valued and thus in an infinite space is appropriate.

Rather than assuming finite bounds and concluding that phonetics-phonology maps are learnable *unconditioned on the choice of phonetic parameterization, as long as the parametrization is finite*, we model the speech signal in $\mathbb{R}^n$ to impose structure on the vast hypothesis space for learning phonetics-phonology maps.

References

Assmann, Peter, and Quentin Summerfield. 2004. The perception of speech under adverse conditions. In *Springer handbook of auditory research*, volume 18, 231–308. New York: Springer-Verlag.

Bard, Ellen Gurman, Dan Robertson, and Antonella Sorace. 1996. Magnitude estimation of linguistic acceptability. *Language* 72:32–68.

Bashford, James A., Keri R. Riener, and Richard M. Warren. 1992. Increasing the intelligibility of speech through multiple phonemic restorations. *Perception and psychophysics* 51:211–217.

Blum, Lenore. 2004. Computing over the reals: where Turing meets Newton. *Notices of the American Mathematical Society* 51:1024–1034.

Blum, Lenore, Felipe Cucker, Michael Shub, and Steve Smale. 1997. *Complexity and real computation*. Springer.

de Boer, Bart, and Patricia K. Kuhl. 2003. Investigating the role of infant-directed speech with a computer model. *Acoustics Research Letters Online* 4:129–134.

Boersma, Paul, and David Weenink. 2010. Praat: doing phonetics by computer (version 5.1.32) [computer program]. `http://www.praat.org`.

Chao, Yuen-Ren. 1930. A system of tone-letters. *Le Maître Phonétique* 45:24–27.

Chao, Yuen Ren. 1968. *A grammar of spoken Chinese*. Berkeley, CA: University of California Press.

Crammer, Koby, and Yoram Singer. 2001. On the algorithmic implementation of multiclass kernel-based vector machines. *Journal of machine learning research* 2:265–292.

Dillon, Brian, Ewan Dunbar, and William Idsardi. Unpublished. A single stage approach to learning phonological categories: insights from Inuktitut .

Divenyi, Pierre L. 2004. Frequency change velocity and acceleration detector: a bird or a red herring? In *Auditory signal processing: physiology, psychoacoustics, and models*, ed. de Cheveigné A. McAdams S. Pressnitzer, D. and L. Collet, 176–184. Spring-Verlag.

Ọdẹ́lọbí, Ọdẹ́túnjí Àjàdí. 2008. Recognition of tones in Yorùbá speech: Experiments with artificial neural networks. In *Speech, audio, image and biomedical signal processing using neural networks*, 23–47. Berlin: Springer.

Dooley, Gary J., and Brian C. J. Moore. 1988. Duration discrimination of steady and gliding tones: A new method for estimating sensitivity to rate of change. *The Journal of the Acoustical Society of America* 84:1332–1337.

Duda, Richard O., Peter E. Hart, and David G. Stork. 2001. *Pattern classification*. John Wiley & Sons, Inc., 2nd edition.

Dyson, Freeman. 2004. A meeting with Enrico Fermi. *Nature* 427:297.

Fields, Stanley, and Mark Johnston. 2005. Whither model organism research? *Science* 307:1885–1886.

Gauthier, Bruno, Rushen Shi, and Yi Xu. 2007. Learning phonetic categories by tracking movements. *Cognition* 103:80–106.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. *The elements of statistical learning*. Springer, second edition.

Hyman, Larry M. 2010. Do tones have features? *UC Berkely Phonology Lab Annual Report* 1–20.

Jain, Sanjay, Daniel Osherson, James S. Royer, and Arun Sharma. 1999. *Systems that learn: An introduction to learning theory (second edition)*. Cambridge, Massachusetts: MIT Press.

Jansen, Aren. 2008. Geometric and landmark-based approaches to speech representation and recognition. Doctoral Dissertation, The University of Chicago.

Kuhl, Patricia K. 2004. Early language acquisition: cracking the speech code. *Nat Rev Neurosci* 5:831–843.

Kuhl, Patricia K, Erica Stevens, Akiko Hayashi, Toshisada Deguchi, Shigeru Kiritani, and Paul Iverson. 2006. Infants show a facilitation effect for native language phonetic perception between 6 and 12 months. *Developmental Science* 9:F13–F21.

Kuhl, Patricia K., Feng-Ming Tsao, and Huei-Mei Liu. 2003. Foreign-language experience in infancy: Effects of short-term exposure and social interaction on phonetic learning. *Proceedings of the National Academy of Sciences of the United States of America* 100:9096–9101.

Ladefoged, Peter, and D. E. Broadbent. 1957. Information conveyed by vowels. *The Journal of the Acoustical Society of America* 29:98–104.

Lam, Hiu Wai, and Kristine M. Yu. 2010. The role of creaky voice in Cantonese tonal perception. In *159th Meeting of Acoustical Society of America, April 2010*.

Lin, Ying. 2005. Learning features and segments from waveforms: a statistical model of early phonological acquisition. Doctoral Dissertation, University of California Los Angeles.

Lin, Ying, and Jeff Mielke. 2008. Discovering place and manner features: What can be learned from acoustic and articulatory data. *University of Pennsylvania Working Papers in Linguistics* 14.

Liu, Huei-Mei, Patricia K. Kuhl, and Feng-Ming Tsao. 2003. An association between mothers' speech clarity and infants' speech discrimination skills. *Developmental Science* 6:F1–F10.

MacKay, David. 2003. *Information theory, pattern recognition and neural networks*. Cambridge University Press.

Maddieson, Ian. 1977. Universals of tone. In *Universals of human language: Volume 2 phonology*, ed. Joseph H. Greenberg, Charles A. Ferguson, and Edith A. Moravcsik. Stanford University Press.

Maddieson, Ian. 1978. The frequency of tones. *UCLA Working Papers in Phonetics* 41:43–52.

Mattock, Karen, Monika Molnar, Linda Polka, and Denis Burnham. 2008. The developmental course of lexical tone perception in the first year of life. *Cognition* 106:1367–1381.

Meyer, A.R., and M.J. Fischer. 1971. Economy of description by automata, grammars, and formal systems. In *12th Annual IEEE Symposium on Switching and Automata THeory*, 188–191.

Mielke, Jeff. 2008. *The emergence of distinctive features*. Oxford University Press.

Miller, George A., and J. C. R. Licklider. 1950. The intelligibility of interrupted speech. *The Journal of the Acoustical Society of America* 22:167–173.

Minsky, Marvin, and Seymour Papert. 1971. Progress report on artificial intelligence. http://web.media.mit.edu/ minsky/papers/PR1971.html.

Narayan, Chandan R., Janet F. Werker, and Patrice Speeter Beddor. 2010. The interaction between acoustic salience and language experience in developmental speech perception: Evidence from nasal place discrimination. *Developmental Science* 13:407–420.

Nearey, Terrance M. 1989. Static, dynamic, and relational properties in vowel perception. *The Journal of the Acoustical Society of America* 85:2088–2113.

Peperkamp, Sharon, Rozenn Le Calvez, Jean-Pierre Nadal, and Emmanuel Dupoux. 2006. The acquisition of allophonic rules: Statistical learning with linguistic constraints. *Cognition* 101:B31–B41.

Peterson, Gordon E., and Harold L. Barney. 1952. Control methods used in a study of the vowels. *The Journal of the Acoustical Society of America* 24:175–184.

Pierrehumbert, Janet. 2003a. Probabilistic phonology: Discrimination and robustness. In *Probability theory in linguistics*, ed. Rens Bod, Jennifer Hay, and Stefanie Jannedy, 177–228. The MIT Press.

Pierrehumbert, Janet B. 1990. Phonological and phonetic representation. *Journal of Phonetics* 375–394.

Pierrehumbert, Janet B. 2003b. Phonetic diversity, statistical learning, and acquisition of phonology. *Language and Speech* 46:115–154.

Pietra, Stephen Della, Vincent Della Pietra, and John Lafferty. 1997. Inducing features of random fields. *IEEE Transactions Pattern Analysis and Machine Intelligence* 19:1–13.

Pike, Kenneth L. 1964. *Tone languages*. University of Michigan, Ann Arbor.

Polka, Linda, Peter W. Jusczyk, and Susan Rvachew. 1995. Methods for studying speech perception in infants and children. In *??????*, 49–89. ??/.

Polka, Linda, and Janet F. Werker. 1994. Developmental changes in perception of nonnative vowel contrasts. *Journal of Experimental Psychology: Human Perception and Performance* 20:421–435.

Qian, Yao, Tan Lee, and Frank K. Soong. 2007. Tone recognition in continuous cantonese speech using supratone models. *The Journal of the Acoustical Society of America* 121:2936–2945.

R Development Core Team. 2010. R: A language and environment for statistical computing. `http://www.R-project.org` ISBN 3-900051-07-0, Vienna, Austria.

Shi, Rushen. in press. Contextual variability and infants' perception of tonal categories. *Chinese Journal of Phonetics* .

Shue, Yen-Liang, Patricia Keating, and Chad Vicenik. 2009. VOICESAUCE: a program for voice analysis. *The Journal of the Acoustical Society of America* 126:2221.

Sundara, Megha, Linda Polka, and Fred Genesee. 2006. Language-experience facilitates discrimination of /d-/ in monolingual and bilingual acquisition of english. *Cognition* 100:369–388.

Tees, Richard C., and Janet F. Werker. 1984. Perceptual flexibility: maintenance or recovery of the ability to discriminate non-native speech sounds. *Canadian Journal of Psychology* 38:579–590.

Tian, Ye, Jian-Lai Zhou, Min Chu, and E. Chang. 2004. Tone recognition with fractionized models and outlined features. In *Acoustics, Speech, and Signal Processing, 2004. Proceedings. (ICASSP '04). IEEE International Conference on*, volume 1, I–105–8 vol.1.

Toscano, Joseph C., and Bob McMurray. 2010. Cue integration with categories: Weighting acoustic cues in speech using unsupervised learning and distributional statistics. *Cognitive Science* 34:434—464.

Vallabha, Gautam K., James L. McClelland, Ferran Pons, Janet F. Werker, and Shigeaki Amano. 2007. Unsupervised learning of vowel categories from infant-directed speech. *Proceedings of the National Academy of Sciences* 104:13273–13278.

Venables, W. N., and B. D. Ripley. 2002. *Modern applied statistics with S*. Springer, fourth edition.

Wahba, Grace. 2002. Soft and hard classification by reproducing kernel hilbert space methods. *Proceedings of the National Academy of Sciences of the United States of America* 99:16524 –16530.

Wang, Siwei, and Gina-Anne Levow. 2008. Mandarin Chinese tone nucleus detection with landmarks. In *Proceedings of Interspeech 2008*, 1101–1104.

van de Weijer, Joost. 1998. Language input for word discovery. Doctoral Dissertation, Katholieke Universiteit Nijmegen, Nijmegen, The Netherlands.

van de Weijer, Joost. 2002. How much does an infant hear in a day? In *Proceedings of the GALA2001 Conference on Language Acquisition*, 279–282.

Welmers, Wm. E. 1973. *African language structures*. University of California Press.

Werker, Janet F., Rushen Shi, Renee Desjardins, Judith E. Pegg, and Linda Polka. 1998. Three methods for testing infant speech perception. In *Perceptual development: visual, auditory, and speech perception in infancy*, ed. A. M. Slater, 389–420. UCL Press.

Werker, Janet F., and Richard C. Tees. 1984. Cross-language speech perception: evidence for perceptual reorganization during the first year of life. *Infant Behavior and Development* 7:49–63.

Widdows, Dominic. 2004. *Geometry and meaning*. CSLI.

Wong, Patrick C. M., and Randy L. Diehl. 2003. Perceptual normalization for inter- and intratalker variation in cantonese level tones. *Journal of Speech, Language & Hearing Research* 46:413–421.

Xu, Nan, and Denis Burnham. submitted. Tone hyperarticulation in Cantonese infant-directed speech. *Developmental Science* .

Xu, Yi. 1997. Contextual tonal variations in Mandarin. *Journal of Phonetics* 25:61–83.

Young, Steve, Gunnar Evermann, Mark Gales, Thomas Hain, Dan Kershaw, Xunying Liu, Garethm Moore, Julian Odell, Dave Ollason, Dan Povey, Valtcho Valtchev, and Phil Woodland. 2009. *The HTK book*. Cambridge University Engineering Department, 3.4 edition.

Yu, Kristine M. 2010. Laryngealization and features for Chinese tonal recognition. In *INTERSPEECH-2010*.

Zhang, Jinsong, and Keikichi Hirose. 2004. Tone nucleus modeling for chinese lexical tone recognition. *Speech Communication* 42:447–466.

Zhou, Ning, Wenle Zhang, Chao-Yang Lee, and Li Xu. 2008. Lexical tone recognition with an artificial neural network. *Ear and hearing* 29:326–335. PMC2562432.

Affiliation

Kristine M. Yu
Department of Linguistics
University of California, Los Angeles
krisyu@ucla.edu