Prosodic Theory and Practice

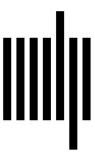
Edited by: Jonathan Barnes, Stefanie Shattuck-Hufnagel

Citation:

Prosodic Theory and Practice
Edited by: Jonathan Barnes, Stefanie Shattuck-Hufnagel
DOI: 10.7551/mitpress/10413.001.0001
ISBN (electronic): 9780262543194
Publisher: The MIT Press
Published: 2022

OA Funding Provided By:

OA Funding from MIT Press Direct to Open



The MIT Press

Introduction: What Are Theories of Prosody For?

Jonathan Barnes and Stefanie Shattuck-Hufnagel

It is one thing to criticize a scientific model on the grounds that the account it provides of some phenomenon seems incorrect or incomplete. It is another thing entirely to criticize a model for never intending to offer an account of something in the first place. Often enough, reasonable scholars just disagree as to the best approach to some phenomenon, and sometimes, at least, this disagreement leads to debate that is productive and progress-oriented. Sadly, this is not always the outcome, and it is our suspicion, furthermore, that at least much of the time in prosody research, the fault for this lies primarily in a misunderstanding of the nature of the criticism being offered. Researchers, we have noticed, often use shared terminology in subtly different ways. These differences, frequently stemming from underlying differences in the goals researchers have set for themselves and their models, are also often either unacknowledged or unexplored by researchers, whose subsequent disagreements are animated precisely by these unspoken divergences.

This volume has two complementary goals, which together we hope can help to address this problem. First, we hope to provide a comprehensive overview of each of the major theoretical approaches to spoken language prosody that are currently practiced, in some cases along with critical commentaries, and responses from the original authors. Second, to deal with the aforementioned "failures to communicate" we fear we witness all too frequently in the world of speech and language research, we felt it was important to ensure that all contributions to the volume address a common set of questions that we believe are critical to a comparative evaluation of what each approach surveyed does best, and what it leaves aside as not central to its enterprise. To this end, we asked that all contributing authors give consideration to a small set of issues that we feel are not currently addressed explicitly, or in the same terms, by all approaches, and yet are generally agreed to be of central importance to the field at large. Our ultimate goal was to elicit the clearest possible treatment in each chapter of two basic, overarching questions: (i) What does the model take to be the central goal(s) of a prosodic theory? and (ii) What are the model's central assumptions (and what would it take to falsify them)? To this end, the specific issues we asked each author to address in one way or another explicitly were these:

1. *Phonology*: What is the role of phonological representations in the model? Does the model see a role for a phonological level of representation in prosodic theory, and if so, what does it look like? Does it posit a set of abstract, symbolic primitives that combine to generate well-formed intonation contours linked both to function/meaning and to the range of acceptable phonetic instantiations?
2. *Meaning*: In what way does the model connect prosodic forms to meanings? For example, is a given intonation contour (taken holistically) linked directly to a given meaning

or function? Or are the meanings of contours derived compositionally from structures built out of meaningful subconstituents? Another way of putting this might be to ask where, if anywhere, is the notion of the morpheme to be located in prosodic structure?

3. *Phonetics*: What is the relation of the model to phonetic implementation? Does the model come with an explicit theory of phonetic implementation, possibly one that could serve as the basis for synthesis of prosody? What role, if any, do the theory's proponents take synthesis to have in the evaluation of competing models of the grammar of prosody?
4. *Typology*: What is the relation of the model to prosodic typology? Does the model make predictions about the kinds of prosodic systems that should, or should not, be found in the languages of the world? What are these predictions?
5. *Psychological status*: What is the psychological status of the model? Is the model imbued with some form of psychological reality? Does it aspire to model cognition, or does it focus instead on successfully modeling or deriving the F0 contours themselves? (Or does it do both?)
6. *Transcription*: What is the relation of the model to systems of prosodic transcription? Is the model useful as a tool for prosodic transcription? If so, is it only a transcription system? If not, is it entirely without implications for transcription?

It is our hope that such an overview of each existing model, addressing at the same time a fixed set of questions identified by the editors, will provide both a useful introduction, and a handy reference guide, for a diverse audience of readers. The fact that these chapters are composed in a parallel fashion, and devoted directly to the same set of questions, will allow readers to compare and contrast each theoretical approach explicitly with all its competitors, in a way that has not to date been convenient. We also hope that the critical commentaries, provided for chapters where we detect in the literature particular levels of ongoing controversy, together with authors' responses, will provide readers with a flavor of current debate in the field, as well as an introduction to the issues most under contention today. Lastly, we hope that this direct comparison of theories on so broad a range of issues as phonological representation, phonetic implementation, prosodic typology, transcription, and corpus management, and so on, will throw into the sharpest possible relief just what is currently known about these matters, and more importantly, what remains to be discovered.

In the following sections, we will review the six questions and their motivations, with the goals of explaining why we felt that specifically these issues were the ones most in need of explicit treatment by the proponents of competing theories of prosody, and also to make maximally clear what we ourselves take the terms invoked to mean. Our aim in so doing is not to present our own views of what a theory of prosody should look like but, rather, to sketch out the range of possibilities for what one might look like, and what particular choices might mean for a given model's most fruitful application. Throughout, we give examples from particular theories, including those presented in this volume, where they most serve to clarify. But in general, we do not attempt to summarize the content of those chapters here, allowing each chapter to speak for itself. Instead, to exemplify what candidate answers to our questions might look like in the context of a single, concrete model, we focus here on one important model that is not presented independently in this volume: the Fujisaki model (Fujisaki and Hirose 1984, and subsequent publications from Fujisaki and collaborators). We have chosen the Fujisaki model for this purpose not because we take any particular stance on the correctness of its approach to the issues, but rather because, correct or

not, this model's developers and practitioners have, over the years, been exceedingly and laudably clear as to what the model's aims and assumptions are, leaving comparatively little to the potential user's imagination as to where it stands on the issues we wish to address. We therefore begin this review with question 1, concerning the nature and role of phonological representations in each model.

I.1 Question 1: Phonology

Ladd ([1996] 2008, especially chapter 1) argues at length that there is a fundamental division to be made between models of intonation that are phonological in nature, and models of intonation that are not. A *phonological model*, to borrow from Pierrehumbert (see commentary in chapter 11, this volume), is one that posits a relatively small set of abstract sound categories that serve as a narrow bridge between the "extremely rich, high-dimensional world of meanings and communicative functions" and the similarly rich and nuanced universe of the acoustic signal. Such a system of sound categories would be effective in managing the flow of information between these two domains both in language learning and in daily use, in that it would allow "articulatory and perceptual patterns exhibited in one word to be reused in other words."[1] In such a system, the encoding and expression of intonation would look more or less exactly like that of segmental phonology, in the sense that the link between a given intonational category and the various meanings it plays a role in expressing would be arbitrary. That is, there should be nothing any more "declarative" about the Autosegmental-Metrical (AM) string H* L-L% in English than there is feline about the string [k^hæt] for *cat*. Furthermore, each instance of a given sound category should be subject to the same set of conditions or procedures for context-specific realization as every other instance of that category, regardless of the meaning or function it serves in a particular construction. All the allophonic variability, patterns of reduction, and so on, that might affect the production or perception of a word-initial /p/, other contextual factors held equal, should apply equivalently, whether the lexical item in question is "potato" or "Parisian." (On this view, what has sometimes been called word-specific phonetics might be usefully viewed as a form of variation dependent on detailed contextual information that includes, for example, predictability.) A *non-phonological model*, by contrast, would be one in which no such inventory of categories existed. Instead, meanings or functions would map in one way or another directly to the acoustics of the signal.

The Fujisaki model is perhaps best thought of as a model of the mapping between phonological representations, about the precise nature of which it is largely agnostic, and the acoustic signal. Fujisaki and Hirose (1984) characterize it as a model of the "control process" of the fundamental frequency. It is a quantitative model, aspiring to accurate, nuanced description of the F0 contour, but crucially, in a manner that reflects the relationship of the control mechanism to the discrete linguistic categories that drive it. While the model took as its original impetus a particular linguistic analysis of Japanese prosody, it has since been extended to model F0 patterns in other languages as well, and, in principle, would be compatible with a variety of phonological analyses of the languages to which it has been applied. Mixdorff and Fujisaki (2000), for example, explore the use of Fujisaki model parameters as a means of mapping between the acoustic signal and a symbolic representation of the kind embodied for German by the AM **To**nes and **B**reak **I**ndices (G-ToBI) system.

The Fujisaki model may not be directly committed to a particular model of phonology, but it does have implications for the structure of any such model. These

implications have at times drawn criticism from practitioners of the AM approach (e.g., Ladd [1996] 2008), and while there certainly are differences between the assumptions of the Fujisaki model about linguistic prosody and the structures normally assumed by AM theorists, a closer comparison also reveals greater commonality than is usually assumed. To see this, though, it is first necessary to explore in a bit more detail the structure of the Fujisaki model.

The bedrock structural principle underlying the Fujisaki model is that the F0 control mechanism involves the superposition of information flowing from two distinct channels. The first of these represents the global shape of the F0 contour over a relatively long, linguistically demarcatable domain, and is called a *phrase command*, while the second represents local pitch events, realized on a shorter timescale, which are called *accent commands*. Events in both channels modulate a "base frequency" taken as input, and the superposition of their effects on that base is meant to yield the specific shape of any given F0 contour. The much-reproduced diagram in figure I.1 depicts the basic structure of the model.

In the original analysis of Japanese, the accent commands were meant to implement the lexical pitch accents of that language, while the phrase commands were meant to represent downtrend (in practice, conflating what might otherwise be distinguished as global declination and local downstep). Because of the nature of Japanese prosody, the original model required neither globally upward-moving phrase components nor multiple accent commands of contrasting types or compositions. In practice, however, nothing would prevent the model from adopting such innovations, as indeed it has, in application to other languages. To model lexical tone in Mandarin, for example, as shown in figure I.2, syllables can be associated with a positive tone or accent command (tone 1, high), a sequence of negative and positive commands (tone 2, rise), a single negative command (tone 3, low), or a sequence of positive and negative commands (tone 4, fall) (Fujisaki et al. 2005). Similar enrichments can yield the contrasting pitch-accent types familiar from the tone and intonation systems of various European languages to which the model has also been applied.

It is tempting to see in the negative and positive accent or tone commands of the Fujisaki model direct analogues of the phonological Lows and Highs that make up representations in the autosegmental tradition, and indeed, in certain analyses, such as the Mandarin one in figure I.2, some parallelism does emerge (H versus LH versus L versus HL). In other cases, however, such as the standard autosegmental analysis of certain English or German pitch accents, the Fujisaki model posits only a positive accent

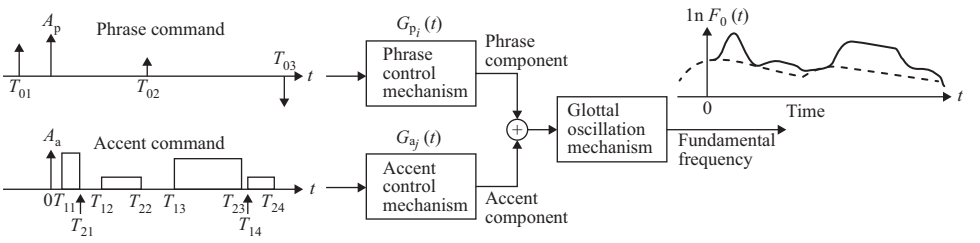


Figure I.1
A functional model for generating F0 contours of sentences. *Source*: Courtesy of Fujisaki (1997).

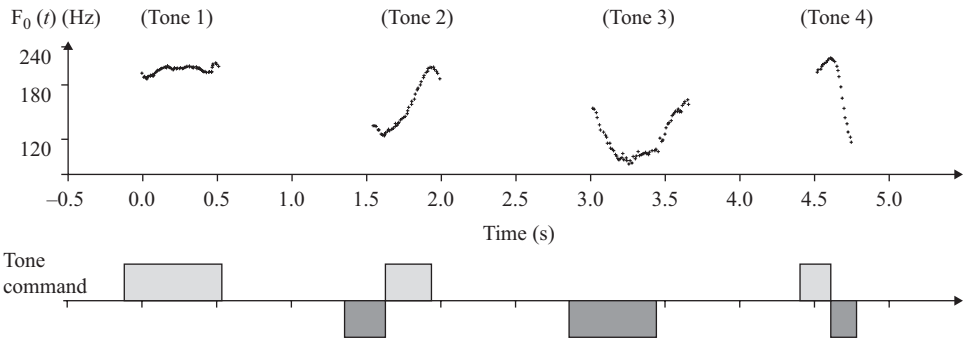


Figure I.2
Fundamental frequency contours for the four Mandarin lexical tone patterns (top) and the underlying tone command patterns used to model them (bottom). *Source*: Courtesy of Fujisaki et al. (2005).

command, where AM models need both a Low and a High to capture the sharp local rise associated with the accent (Fujisaki, Ohno, and Wang 1998). Where the parallelism breaks down even more obviously, though, is with the Fujisakian phrase command, which has no direct analogue in autosegmental terms. The phrase command is commonly understood, even by its proponents, to interpret a "global" component of the F0 contour. Therefore the phrase command appears to be something antithetical to the spirit of the standard AM approach, which purports to view tonal representations of all kinds, whether associated with prominences (e.g., pitch accents) or phrase boundaries (e.g., boundary tones or "phrase accents"), as equivalently just a sequence of locally instantiated tonal targets, instructions to reach a specific F0 level by a specified point in time.

But if Fujisaki's phrase commands are global in the sense that they determine the characteristics of the F0 contour over a relatively longer duration, they are not global in the sense that they directly encode a target shape for the F0 contour over a specific duration (as do the "grids" posited by Gårding 1983, or the "superordinate intonation contours" described by Grønnum in chapter 2, this volume). Rather, both the phrase and accent components of the model are implemented as strings of locally deployed "neuromotor commands." The effects of the phrase component on the F0 contour are modeled as the response of a critically damped second-order linear system to an impulse with a particular timing and amplitude. The longer timescale of these events in comparison with that of the accent commands is due to the response characteristics of the system, as it decays back to a specified "base frequency." The phrase impulse itself, however, is not extended in time. In some ways, it is Fujisaki's accent command that explicitly imposes particular F0 characteristics over a temporally extended domain, insofar as it is modeled as a step function that requires for its implementation explicit times of onset and offset. This is especially clear in figure I.3, where the phrase and accent components for the analysis of a sample utterance of Japanese are depicted separately.[2]

If it isn't globality per se, then, that makes AM and the Fujisaki model incompatible, perhaps it is some other aspect of the notion of *superposition*, or the division of the control mechanism into multiple distinct levels. But this too doesn't obviously divide the approaches. Superposition in the Fujisaki model does mean that the F0 at any particular moment in an utterance can represent the interaction of commands from

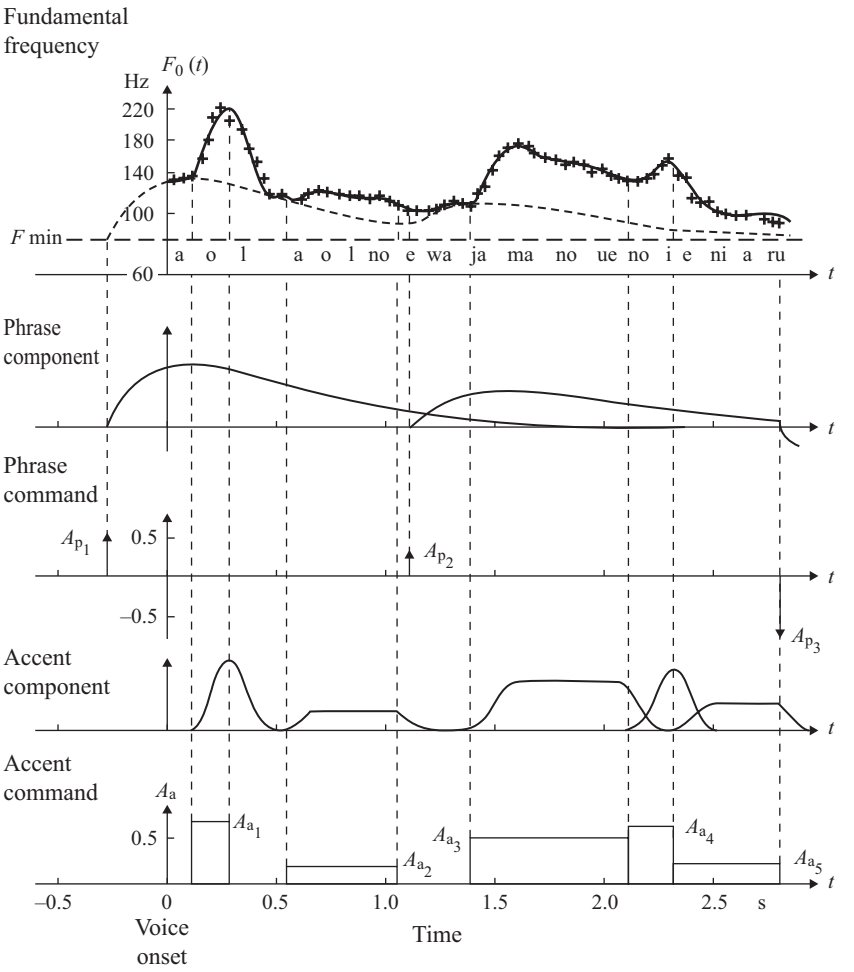


Figure I.3
Decomposition of the F0 contour for the Japanese sentence /aoi aoi no e wa jama no ue no ie ni aru/, 青い葵の絵は 山の上の家にある (The picture of a blue hollyhock is in a house on the top of the mountain) into phrase and accent components, along with the underlying commands that give rise to them. *Source*: Courtesy of Fujisaki (1997).

several distinct representational objects (rather than the apparent sequential movement from one discrete target to the next that ostensibly characterizes AM phonetic interpretation). But this is hardly an unusual situation in the phonetics-phonology interface. Indeed, even in pre-autosegmental days, when phonological representations consisted more or less exclusively of strings of symbols in strict linear arrangement with respect to one another (as per, e.g., Chomsky and Halle 1968), it seems to have been accepted as unproblematic that the state of the vocal apparatus at any given moment should simultaneously reflect the influence of multiple elements of those strings. The symbolic representations of consonant and vowel features relating to place of articulation may not overlap with one another, but measured F2 values during a

given vowel, for example, will simultaneously reflect a number of factors, including "inherent" targets projected by the vowel itself, coarticulation with neighboring vowels and neighboring consonants, modulations caused by prosodic factors related to timing and prominence, and so forth.

Autosegmental representation of phonological features only amplifies this possibility, in that features share hosts (e.g., timing slots, root nodes) in ways that blur the edges of the segment as traditionally understood. Indeed, the whole notion of a tonal tier that exists in parallel to the representation of the segmental melody is nothing if not superpositional, in the sense that both the tonal targets, and the various laryngeal states demanded by the consonants and vowels of an utterance, must ultimately be realized simultaneously, by a single larynx. In some ways, then, perhaps the Fujisaki model is more autosegmental in spirit than AM itself. Perhaps the phrase and accent components could fruitfully be understood as analogous to autosegmental tiers, each housing distinct, but related kinds of pitch specifications. (C-place and V-place features come to mind [Clements 1991], as do, closer to home, distinct register and tone tiers [Yip 1980].) Ultimately, in both models distinct streams of information must eventually be conflated, yielding the temporal overlap of multiple distinct information sources that is typical of phonetic realization more generally.[3]

Regarding the need to posit two distinct types of pitch specifications in the Fujisaki model, one reason that AM models are able to achieve the appearance of representational uniformity is just that they effectively fail to provide any uniform treatment of the phenomena that ultimately motivated the phrase component of the Fujisaki model in the first place. Downtrend, reset, and various other phenomena involving baselines and register have been treated in various ways by different AM theorists, but they are more often than not simply absent from the symbolic representations of intonation contours in those models, with the understanding that they must arise somehow phonetically, through the interpretation of prosodic phrase structure, metrical relations, tonotactics, or some combination thereof. Even excluding these phenomena from consideration, though, AM representations of prosody are not always interpreted as uniformly as the target-and-interpolation approach suggests they should be. While pitch accents and certain boundary tones sometimes seem to project unique, localizable phonetic targets (identified with "turning points" in the F0 contour), the phrase accents of some analyses (Pierrehumbert 1980; Grice et al. 2000) seem to be governed by realizational principles of a different stripe, involving targets extended over broader temporal domains. While various explanations have been offered for this behavior (among them, autosegmental spreading, copying, secondary association), it still seems odd that it is this one category of tonal event that is so consistently in need of such devices. None of this is to suggest that either model is "correct" in its assumptions, but rather we are pointing out that AM models may not be so representationally uniform, nor Fujisakian superposition so alien to them, as is commonly assumed.

In sum, a phonologically based model of prosody will include a set of abstract symbolic sound categories, themselves without meaning, that can be combined into strings subject to the restrictions of the grammar, and that serve as a narrow bridge between the rich universe of meanings and the similarly rich universe of sound shapes. In contrast, a phonology-free model may map meanings directly onto sounds, without an intervening set of abstract categories, or it may simply remain agnostic as to the nature of the categories involved in phonetic implementation.

I.2 Question 2: Meaning

The second question we posed to contributors concerned the place of meaning in their prosodic models. We suggested in the formulation of the question that what we really wanted to know was where, if anywhere, the notion of the *morpheme*, or minimal meaning-bearing element, resides in their approaches. Using question 1 as a springboard here, if there is such a thing as intonational phonology, and if that thing is similar in nature to segmental phonology, then we must expect the tonal primitives of the system to be themselves devoid of meaning, just as individual consonant and vowel phonemes are said to be in non-sound-symbolic vocabulary in the classical Saussurean sense. Those phonological primitives only are associated with meanings once they are combined, according to the constraints imposed by the phonological grammar, into strings of varying lengths, which in spoken languages become the signifiers that join together in arbitrary but inviolable union with their signifieds, to make up the linguistic sign. This duality of patterning (Hockett 1958; Ladd 2014) or double articulation (Martinet 1949) has been said to be fundamental to the structure of human language in general, and is one of the more frequently mentioned properties that may (or may not) separate human language from other animal communication systems. (E.g., Cheney and Seyfarth's [1990] vervet monkeys could be argued to have mastered arbitrariness, but apparently still lack double articulation.) It is this "design feature" of language that allows the creation of the large, potentially infinite vocabularies characteristic of human languages out of small, finite sets of phonological elements (cf. Sandler et al. 2011 on the emergence of this property in linguistic systems, such as new sign languages, that may not originally have it).

Speaking only of intonation, it is perhaps not clear whether such a feature is necessary to the expression of the meanings encoded by the system, in part because it is still not clear how numerous, or even how enumerable, the meanings in question are. Still, it is somewhat surprising that in the AM tradition, at least in one of the most influential approaches, that of Pierrehumbert and Hirschberg (1990), the traditional roles occupied by the phoneme and morpheme appear to be conflated. In that approach, the individual tones making up the tone string (e.g., H*, L-, L%), potentially down to the component targets of bitonal pitch accents (e.g., the leading L of L+H* or the trailing H of L*+H [301]) are associated each with their own elements of meaning. Meanings conveyed by entire utterances are then argued to be built up compositionally from those elements, rather like sentence meanings are composed from the combinations of the morphemes they contain into particular structural configurations.[4] But if individual tonal targets are effectively morpheme-like, they are not themselves obviously composed of any phoneme-like subparts, which would appear to make intonational phonology quite different from the rest of phonology, in that it does not seem to show duality of patterning. Some of these issues are discussed interestingly by Liu et al. (2013), where it is argued that "prosodic functions," such as focus and modality, are the morpheme-like elements in intonational meaning and that their formal expression must therefore be something more akin to the whole contour than the individual tonal targets (see also Xu, Prom-on, and Liu, chapter 11, this volume, for an exposition of the model assumed in that study).

To continue giving sample answers to our questions using the Fujisaki model illustratively, here the Fujisaki model has no particular commitments that we are aware of. As a model of phonetic implementation, it assumes as its input some form of "linguistic information," which Fujisaki (1997: 28) describes as "symbolic information that

is represented by a set of discrete symbols and rules for their combination." (In this instance, he has in mind the accent types of Japanese words, but one could easily substitute Mandarin tones, English pitch accents, or any other F0-based linguistic entities.) It is fair to say that capturing how these entities relate to the expression of meaning is not among the central goals of the model.

In sum, two key questions about how a prosodic model relates to meaning are whether the meaning-form mapping includes an intermediate symbolic representation in the shape of elements that themselves do not bear meaning, and whether the meaning-bearing elements are individual tonal targets, strings of targets, or some other type of unit.

I.3 Question 3: Phonetic Implementation and Synthesis

Phonetic implementation is perhaps the one area where all the models included in this volume have some explicit position or stake, and thus it might appear to be the single area most promising for direct "head-to-head" comparison or evaluation of competing models of prosody. Even in this area, though, differing aims and different assumptions make the relative "success" of the models less directly evaluable than is commonly assumed.

Once again (and if this book as a whole could be assigned a single take-home message, this would be a strong candidate), which model of prosody is seen as "correct" or "best" depends on what the user wants to do with it. No one model may answer all possible needs. For example, a model that aims to understand prosody in the context of human cognitive architecture might also yield insights about cross-language typology of prosodic systems, but then again, it might not. (The argument concerning the sources of cross-language patterns in phonology—that is, whether they are best thought of as directly encoded in grammar or linguistic cognition, or whether they might instead arise as by-products of the transmission process, from facts about articulation, speech acoustics, domain-general properties of the auditory systems, and so on—is presumably as relevant in prosody as it is in segmental phonology.) The model that currently yields the most lifelike synthesis might similarly have little to tell us about cognition, and so forth. Convergence in these domains of inquiry is possible, and even perhaps desirable, but it is not logically necessary.

The Fujisaki model is particularly rewarding to examine in this connection, because its aims in this regard are both exceptionally clearly stated in a host of publications (e.g., Fujisaki and Nagashima 1969; Fujisaki and Hirose 1984; and Fujisaki 2004, among others), and interestingly divergent from the aims of some of the models included in this volume, even where they might initially appear to be quite similar. The Fujisaki model was "conceived and developed primarily for use in (high-quality) speech synthesis" and was not originally concerned with capturing the cognitive aspects of language prosody. At the same time, it is clear that from its earliest incarnation it had the ambition not just to produce realistic-sounding synthetic speech for technological applications, but also to provide a tool by which we might learn things about the properties of human languages. Both the architecture and implementational details of the model are "deeply based on the physiology and physics of the mechanisms and processes of controlling the frequency of vibration of the vocal folds." It therefore "models a part of the process of speech production" (Hiroya Fujisaki, personal communication, 2016).

It is worth pausing at this juncture for a moment to appreciate how truly, deeply unnecessary this characteristic is, if our primary aim is nothing more than to allow

electronic devices to produce synthetic utterances with natural-sounding prosody. The success of the Fujisaki model in producing high-quality synthesis shows us that it is possible for such a model to succeed, but there is no a priori reason, as far as we know, to assume that the best means for production of natural-sounding synthetic speech using a digital computer will require formulae that are conceptually parallel in any way to the control mechanisms for fundamental frequency in the human vocal tract. It is logically possible, for example, that the best synthetic prosody would in fact be produced by a model whose properties had no parallels at all to human physiology, human cognition, common grammatical properties of human languages, or anything else we might find interesting in the realm of prosody. It is entirely possible that the model that proves most useful from a technological point of view will also prove least interesting from a scientific one.

Fujisaki's goal, by contrast, is not just to synthesize natural-sounding Japanese, but in so doing, to learn something about the nature of human language. By structuring the model in such a way as to "quantitatively simulate the control process," we may then be able to use the model to "separate the characteristics of the physical and physiological mechanisms from the factors that carry linguistic information" (Fujisaki and Hirose 1982: 69). To the extent that those physical mechanisms are common to all members of the species, any differences between languages that emerge through this modeling may be assumed to give us insight into structural differences between the languages in question. Speaking more generally of what makes for a successful scientific model, Fujisaki says:

> In my humble definition, a model is a description (preferably in objective, physical, and mathematical terms) of the essential structure of a mechanism (a structural model), or the essential function of a process (a functional model). Thus it is not a subjective statement or description, but is an objective description, most preferably by mathematical terms. Note, however, that it is different from *a mere approximation by some arbitrary mathematical functions*. In my humble personal view, mathematics is a tool to describe fairly precisely what we observe in physics, physiology, or in cognitive psychology, but should be based, not on surface phenomena, but on the underlying mechanism or process that produces the observed phenomena. (personal communication; emphasis added)

In a similar vein, Hans-Jörg Mixdorff argues, concerning the insistence on "linguistically meaningful model parameters" in the Fujisaki model:

> An arbitrary number of commands provided, any F0 contour can be approximated with unlimited accuracy. For this reason constraints must be applied in order to ensure a linguistically meaningful interpretation of the analysis results. These constraints are language specific and concern the relationship between linguistic units and structures (prosodic phrases and accents for instance) and the phrase and accent commands. (1998, 51)

The phrase and accent components of the Fujisaki model are thus not adopted with their particular mathematical characteristics just because they exhibit success in modeling the F0 contours observed in natural speech, but these components are in fact identified with distinct aspects of the physiology of F0 control as well. Specifically, the activity of the phrase component is said to reflect horizontal translation of the thyroid cartilage stemming from the activity of the pars obliqua of the cricothyroid muscle, while the accent component reflects rotation around the cricothyroid joint caused by contraction of the pars recta of the same muscle (Fujisaki 1981, 1988). The differing timescales of the two components of the model, and hence the linguistic phenomena they are selected to implement (global or phrasal versus local or accentual), are thus

argued to follow from the temporal properties inherent in the control of these two degrees of freedom in the movement of the thyroid cartilage relative to the cricoid. Whatever else these aspects of the model might be, then, they are clearly not arbitrary.

Fujisaki's work is also noteworthy in being among the first to invoke the notion of analysis-by-synthesis as a procedure for scientific investigation in the area of prosody. This is unsurprising, because Professor Fujisaki was present and in collaboration with researchers at MIT when the theory of analysis-by-synthesis was being developed. Indeed, he was second author on one of the earliest publications of research involving the theory (Bell et al. 1961). Since that time, outgrowths of the original analysis-by-synthesis concept have become important, and widely cited, in the prosody literature. We have noticed, however, a great deal of diversity in what different researchers seem to mean when they invoke this term today, so much so that, as with other terms (e.g., *phonology*) that have become commonplace in our discourse, we suspect that the comfort of shared terminology may be seducing us into the belief that we understand one another's ideas better than we in fact do.

With regard to the term *analysis-by-synthesis*, the original concept was first published, as far as we know, in a 1959 paper by Halle and Stevens. Described in that paper as a "dynamic system for the analysis of signals," (1) the theory was apparently intended to be used in two distinct but related ways. First, it was proposed as a model of a parsing strategy—of what "receivers" (in spoken language, listeners) may be doing when they decode linguistic signals of various kinds. Second, it was intended as a tool for use in the then quite new field of automatic speech recognition. Early in their paper, Halle and Stevens characterize the speech signal as a "particular value of some function," which they refer to as the "encoding function." The argument of this function they refer to as the "message," and therefore they represent encoding and decoding as involving the following relation (1):

$$\text{Signal} = F(\text{Message})$$

A notable property of most linguistic communication, they continue, is that, because every listener (or decoder) is at least in principle some of the time also a speaker (or encoder), listeners, when faced with the task of decoding a given signal, have access to both the same encoding function and the same basic inventory of possible messages that the speaker has. What Halle and Stevens suggest, therefore, is that the decoding process may involve a kind of "active internal replication process" on the part of the receiver. Faced with a particular spoken signal, in other words, a listener takes a set of candidate intended messages, runs those through the encoding function himself or herself, and compares the resulting internally generated signals to a stored version of the original. The candidate message yielding a signal with the least amount of error relative to the original is assumed to be what the speaker intended to communicate.

The same sort of procedure is proposed as a tool in the automatic analysis of a remarkably broad set of different kinds of linguistic signals (e.g., vowel formants, handwriting, Russian morphology). For example, Bell et al. (1961) propose a method to use analysis-by-synthesis as a step toward the automatic recognition of non-nasal vowels, by using the method to figure out the vocal tract resonances and source characteristics that were used to produce a particular spectral slice taken from the signal. As the authors of that paper explain, however, this is possible only because of the explicit models of the encoding function that we have, in the form of the acoustic theory of speech production (Fant 1960). The Signal to be decoded is the vowel spectrum

sampled at a particular point in time, and the Message is "a tabulation of the resonant frequencies and relevant data concerning the excitation." (3) Further, Halle and Stevens (1959) propose that this same procedure could be used in the "second stage" of the work of a speech analyzer, this time taking those same resonant frequencies and source characteristics to be the signal, from which could be decoded a message consisting of a string of "discrete phonetic symbols." (3)

In yet another application, it is reported that Halle was at work with a colleague on a system for the automatic parsing of the morphological structure of Russian words. Halle and Stevens state explicitly that the reason such a morphological processor is even conceivable is that we believe we have a good understanding both of what the complete set of possible elements in a Message would be (the inventory of Russian morphemes), and also of the encoding function, or rules used by speakers for those morphemes' combination. It should be clear that if we lacked either of those elements, automatic morphological parsing of Russian words would not be possible using the analysis-by-synthesis method.

As with the model of vowel perception, the work of a human listener in morphological processing might proceed analogously to the proposed automated system. Because the listener and the speaker share both an inventory of possible input morphemes and the rules for their combination, a Russian listener confronted with a particular spoken lexical item might think "What set of morphemes, in what structural relationship, would it take to yield a word comparable to the one I just heard?"

In the area of prosody research, analysis-by-synthesis entered researchers' lexicons very early on. By 1965, it appears in the title of a paper by Öhman and Lindqvist ("Analysis-by-Synthesis of Prosodic Pitch Contours"), which does not, for whatever reason, cite Halle and Stevens. (Oddly, the phrase also does not appear in that paper anywhere other than in the title, making their precise conception of it difficult to ascertain in retrospect.) Fujisaki, however, who describes the genesis of his own model in large part as a response to certain aspects of Öhman's model, uses *Analysis-by-Synthesis* (which he tends to capitalize) in more or less the same way that Halle and Stevens (1959) and Bell et al. (1961) used it. Because he has an explicit quantitative model of Japanese prosody, including an encoding function, and a hypothesis about the kinds of elements that serve as its inputs, he describes how analysis-by-synthesis can be used to analyze a given Signal (a raw F0 contour) in terms of the Message (phrase commands, accent commands, and their corresponding parameter settings) that must have given rise to it. In Fujisaki (1997), he also talks about a second stage of analysis, just as Halle and Stevens did. If the first level of analysis infers the phrase and accent commands from the characteristics of the signal, then the second application of analysis-by-synthesis infers the "units and structures of prosody" (30) from the commands. Obviously, this last level of analysis is only possible assuming we have some idea what those units and structures are, and what the encoding function that maps from them to a particular set of phrase and accent commands is like.

There is, however, another common use of the term *analysis-by-synthesis* in the prosody literature, and that is as a method for the validation of scientific models. Daniel Hirst (2011), who has employed the concept of analysis-by-synthesis in his work since at least 1980, summarizes this take on the model neatly:

> The analysis by synthesis paradigm is potentially an attractive one for linguists, since it provides an empirical solution to the problem of validating an abstract model. The interaction between linguists and engineers has always been a productive area of exchange. This

> is particularly evident in the area of speech prosody. If the representation derived from a model can be used as input to a speech synthesis system, and if the contrasts represented in the model are correctly rendered in the synthetic speech, then the representation can be assumed to contain all the information necessary to express that contrast. (58)

This notion is clearly present in the original papers too, though it appears more as an aside than as an end in itself. Bell et al. (1961), for example, remark that "once a set of parameters is found such that a good replica of the input signal is generated when these parameters are applied as instructions to the internal generative model, then there is little question that this set constitutes an adequate representation of the input." (1735) Fujisaki (1997) says of his own model, again in passing, that "the close agreement of the model's output with the measured F0 contour, found in this as well as in a number of speech samples analysed, attest the validity of the model." (35)

Unfortunately, it is all too easy to take this observation a step further, and to infer that if good synthesis validates a model in some sense, then better synthesis indicates a superior model. Of course, within the confines of a single model, with a given encoding function, and a particular set of candidate input elements, analysis-by-synthesis tells us that the analysis most closely reproducing the signal is the best one. One set of phrase and accent commands, for example, and not another, might be the best decoding of a given F0 contour. But beyond this, the notion of model comparison through analysis-by-synthesis breaks down. We might in principle want to decide, for example, which theoretical model of prosody was "better," the AM model or the Fujisaki model, by comparing the quality of synthesis produced using the two. This might be a good idea, or it might not be. Either way, though, it would not constitute an application of the principle of analysis-by-synthesis. The reason should be clear, returning to that basic relation, Signal = F(Message). To analyze a given signal, the AM model and the Fujisaki model (one imagines) would employ completely different encoding functions, as well as entirely different sorts of representational primitives for their "messages."[5] By the technique of analysis-by-synthesis, given a shared encoding function, and a shared inventory of possible messages, we can determine which of two (or more) messages was more likely to have produced the signal in question. Without that common ground, though, no such inferences are possible.

We might still feel, nonetheless, that quality of synthesis is a suitable criterion by which to decide between prosodic theories, even if calling this procedure analysis-by-synthesis is historiographically misleading. Again, in this instance, we would simply invoke the caveat that in order to decide which of two models is "better," it is important first to answer the question: Better for what? Though as we understand it, this may no longer be the case, there was a moment in the history of the field of speech synthesis when the best programs for producing synthetic speech were based on concatenation of units called "diphones." *Diphones* are essentially transitions from one segment, traditionally conceived, to another. Rather than concatenating a series of consonants and vowels, in other words, diphone synthesis would concatenate a unit containing the transition from a particular consonant into a particular vowel, with the transition from that same vowel into the following consonant, and so forth. Despite the successes of diphone synthesis, however, few linguists were moved to posit that perhaps diphones were the correct representational primitives to be used in our phonological models as well. To the extent that diphones did not replace traditional consonant and vowel segments as elements of phonological representation, this was presumably because phonologists felt that whatever was gained by their adoption for purposes of

synthesis was offset by the insights that would be lost concerning the phonological structures of individual languages, cross-language phonological typology, and so forth. Their aims, in other words, were different.[6]

I.4 Question 4: Typology

The purpose of this question is to determine the extent to which the prosodic models included in this volume address questions of "possible" and "impossible" prosodic systems, cross-language frequency of prosodic systems or processes, and so on. For as long as there have been phonologists, there has been a general acknowledgment of responsibility among them to questions of why certain kinds of phonological patterns (vowel inventories, rule interactions, stress systems, and so on) should be common (if not universal), while others should be rare (if not unheard of).

What constitutes an "explanation" for such cross-language patterns is the subject of much discussion in the field, the sum of which cannot be adequately reviewed in this space. (See Hyman 2018a and Gordon 2016, chapter 2, for two illuminating, though quite different, overviews of these issues.) For our purposes, it is enough to note that in some research traditions—for example, most approaches to phonology within the umbrella of Optimality Theory (Prince and Smolensky 1993)—the "typological coverage" that a proposed theoretical innovation achieves is a central, even the central, criterion by which its fitness is evaluated. Both *overgeneration* of language patterns (that is, predicting the existence of unattested grammatical systems) and *undergeneration* (failure to predict attested systems) are considered shortcomings from this point of view (though in practice, overgeneration is usually understood to be a lesser sin than undergeneration, because currently unattested language types could nonetheless turn out to be possible). If typological coverage is a common goal for the proprietors of linguistic theories, though, it is by no means a necessary one. Researchers disagree as to which, if any, cross-language regularities should be encoded in "the grammar," where that term usually, if not always, carries cognitive implications. One place this can be seen is in recent controversy over the relative roles of synchronically versus diachronically oriented explanations for typological patterns. Synchronic explanations commonly invoke what Moreton (2008) calls *analytic bias*, a cognitive predisposition, whether language-specific or domain-general, that makes a particular pattern difficult or impossible to learn or predisposes learners to favor one kind of pattern over another. Diachronic explanations, by contrast, focus on what Moreton called *channel bias*—some aspect, in other words, of language use or transmission that makes phonological changes giving rise to one kind of pattern more likely or more common than others. Synchronic explanations also tend to involve hypotheses about "higher-level" aspects of cognition, focusing on facts about learning strategies or computational complexity, while diachronic explanations tend to prefer "lower-level" or entirely noncognitive approaches, featuring instead facts about speech production, speech acoustics, audition, or perception. Thinking of synchronic explanations as strictly cognitive, and diachronic ones as noncognitive, though, is not quite right. Much of the literature on analytic bias focuses on the possibility that at least some of that bias is "substantive" in nature, meaning that humans are predisposed to learn patterns that are "phonetically natural" (where *natural* is usually taken to mean functionally preferable in light of facts about articulation, audition, and so on). In addition, as Kiparsky (2006) points out, sound change is obviously shaped by a cognitive or grammatical component as well. Opinion currently seems settled, following Moreton (2008), that both analytic

and channel biases have a role to play in shaping typology. There is also some evidence (e.g., Moreton and Pater 2012; Glewwe et al. 2018) against the existence of *substantive bias*, or the preference in language learning for patterns that are phonetically natural. Many key questions, in any case, have yet to be resolved.

Our purpose in discussing this at all is just to point out that it is logically possible to wish to devise a cognitive model of how, for example, prosody works, but to believe that much or all of what we think of as prosodic typology has no place in the development or evaluation of that model. Likewise, it is entirely possible that the most successful approach to the synthesis of natural-sounding prosody will have nothing whatsoever to tell us about typology. (Obviously, in both these cases, the opposite could also be true.) What we wish to emphasize is that complaints that a psycholinguistic model fails to account for typological patterns will be constructively received only if the proponents of the model in question believe it is their responsibility to account for facts about typology in the first place. Complaining that a typologist has failed to deliver up a functioning system for prosodic synthesis may be similarly beside the point. As with the preceding three questions, though, we fear that in many instances, most of us are simply unaware of the commitments of many models in this regard, and prefer instead to hold every model tacitly responsible for whatever interests us personally the most in prosody, whether that forms part of that model's intended purpose or not.

To continue our extended case study, the Fujisaki model is not explicitly, centrally typological, but, as should be obvious from our description of its properties above, it does make clear predictions about what prosodic systems should or should not be like. The combination of phrase and accent commands with the full range of their potential parameter values will generate F0 contours of some shapes, and will not generate others.

Application to a variety of typologically distinct languages has always played a role in the model's development, and it has also at times been criticized (e.g., Ladd [1996] 2008, 28–29) for undergenerating attested F0 contour shapes. We are reasonably certain, furthermore, that if such criticisms proved correct, the models' proponents and developers would view that as a shortcoming, and seek to remedy it. More deeply, though, the very proposal that F0 events come in two distinct control varieties, corresponding to the phrase and accent commands, has ramifications for the predicted typology of human language prosody.

I.5 Question 5: Psychological Reality

At least since Sapir (1933), if not much earlier,[7] phonologists have been drawn, moth-like, toward the beckoning flame of psychological reality. Much of contemporary phonological theory, to the extent that it seeks to model "competence," has at least a pretension to some form of psychological explanation. Some work presents this aim explicitly (e.g., Chomsky and Halle 1968, who from the beginning speak of grammar, their object of investigation, as a "system of rules represented in the mind of the speaker-hearer" [1]; or Prince and Smolensky 1993, chapter 10, who situate Optimality Theory relative to other approaches to the cognitive modeling of language, such as connectionism). Other work, sometimes ostensibly in the same tradition, either downplays the connection to cognition (Hyman 2018b, on "what's in the language" versus "what's in the head" [597]) or actively repudiates it (Ladefoged 2005 on language "as a social institution rather than a mental concept" [1]). Often, though, and the attentive reader will no doubt have begun to recognize a theme here, it is not always clear where a given analysis or model stands in this regard, inviting others to make assumptions

that aren't always warranted. As with the other questions we have addressed, there is little sense in criticizing a model for being cognitively implausible if it makes no pretense to model cognition in the first place. One could always argue that a decent model should in fact attempt to model cognition, but again, this is a complaint of a very different nature.

At its inception, the Fujisaki model was not focused, directly at least, on questions of cognition. At the same time, to the extent that the phrase and accent components are characterized as modeling sets of "neuromotor commands carrying linguistic information" (Fujisaki, Ohno, and Wang 1998, 1), it is clear that these elements are meant to be "meaningful both linguistically and cognitively" (Fujisaki, personal communication).

I.6 Question 6: Transcription

A recurring question in the development of transcription systems, for either segmental or prosodic purposes, is whether to attempt to capture the contrastive categories of the language that underlie the speech signal, or to focus on the wide range of ways in which each category can be appropriately realized. With regard to segmental transcription, it is surely no exaggeration to claim that the creation and propagation of the International Phonetic Alphabet in the late nineteenth and early twentieth centuries laid groundwork without which we might never have seen the rapid progress in phonological thought that followed closely on its heels, and continues to the present day. There is a basic sense in which it is necessary to achieve (or at least approach) consensus as to *what* has happened in the course of some natural phenomenon before we can set about formulating explanations as to *why*, or *how*, and the IPA provides a way of doing that for spoken utterances. Thus, it has practical value to the extent that it provides us with a set of labels that, at least in principle, should be universally applicable and interpretable for the characterization of spoken language data, regardless of the language or context in which the speech was produced, and this was obviously and rightly one of its developers' central goals as well. Perhaps less obviously, and more importantly from a theoretical point of view, it also served to cement, behind those labels, an expanding set of putatively universal phonetic categories that should recur with greater or lesser regularity across the languages of the world. The idea is that we ought to be able to agree, for example, on whether or not a given speaker, in a given utterance, has produced an instance of voiceless, unaspirated, bilabial stop, and that we should be able to do so regardless of the phonological analysis we might like to give to that event down the road. Whether it represents an allophone of the phoneme /b/, or of /p/, or of something else again, depends on the language; the morphophonology of the lexical item in question; and so forth. That it was, regardless, an instance of phonetic [p], however, should in principle be the unimpeachable bedrock upon which any analysis must eventually stand.

The notion of universal phonetic categories has its problems, of course, and its detractors have not been shy in saying so. While parts of this universal inventory have seemed relatively uncontroversial, in other parts it has been less clear whether or where to draw boundaries between would-be categories, and whether to dignify particular candidates with their own symbols, or to relegate them instead to expression via diacritics draped around more respectable categories. Varieties of obstruents occurring somewhere around the alveopalatal (or alveolopalatal, or palato-alveolar, or postalveolar, or prepalatal) places of articulation, with or without concomitant tongue posture specifications, spring to mind.

The principle designed to mitigate the potential for unfettered expansion of the system of phonetic categories is that for a new symbol to be introduced, with all its ontological entailments, the sound in question must be contrastive in some language.[8] This is intended to guard against the adoption of distinct symbols to represent sound characteristics that are judged to be universally a matter of subphonemic detail rather than the sole feature minimally distinguishing one sound from another. In practice, though, it can be difficult in many cases to decide how such a principle should be applied.[9] Often the question demanding an answer seems to be, "Contrastive with what?" For example, one area in which the current state of the official IPA chart suggests a breakdown of application of this principle is in the profusion of variant symbols for degrees of nuance in the mid-central portion of the vowel space. The IPA chart currently contains six distinct symbols (excluding vowels such as [ɨ], [ʊ̈], or [ʌ], which are officially high or back, despite their occasional use as symbols for things lower or fronter than that seems to imply). These are [ɘ, ɵ, ə, ɜ, ɞ, ɐ]. In principle, these are all distinct vowel categories. We know what they are meant to sound like, and if we are ever in doubt, recordings exist of consensus productions to remind us. This is not the place to rehearse the individual careers of all these symbols.[10] No doubt it is the case that for each of them there exists, in some language somewhere, a mid-central vowel whose canonical realization sounds more like that one symbol than it does like any of the others. And presumably, that vowel is "contrastive." But with what? Most vowel systems that we know of have one mid-central vowel, if they have any. Some have two (assuming one of them is not better considered high, or back, or front, or what have you).

Obviously, no system has all of these vowels in contrast with one another. Indeed, it seems unlikely to us that even each potential two-way contrast implied by this set of symbols is convincingly represented in some attested language. But if it feels worrisomely like we have embarked on an effort to populate what is effectively a continuous space with discrete symbols whose optimal, final number is not specified by any existing principle, we might nonetheless take comfort in the thought that, with all these symbols, we're at least not likely to be missing anything.[11] For any given real-world mid-central vowel we might encounter, in other words, chances are we have a symbol that does it justice. Yet even so, the fussier of us remain dissatisfied (e.g., Barnes 2006), feeling compelled to note pedantically, for example, that while the IPA standard transcription for the Bulgarian vowel spelled "ъ" and traditionally romanized with "ă" is [ʊ̈] (Ternes and Vladimirova-Buhtz 1990), in fact the vowel in question sounds lower, and fronter, than that symbol implies, if, however, not so low and so front as to make any other IPA symbol more appropriate than that one. We have achieved, in other words, the worst of both worlds: a system at once with too much and with too little nuance built into it.[12]

This lengthy preamble is only to underscore the seriousness of the fact that, whatever uncertainty we might encounter with the transcription of segmental phonology, the situation with respect to the transcription of prosody is in every respect much, much worse. If there are parts of the vowel space in which natural boundaries between universal phonetic categories seem elusive, many, if not most, dimensions of segmental contrast seem reassuringly stable (at least at first glance). In prosody research, though, our abiding fear is that in the end, it may just be mid-central vowels all the way down. In thinking about the transcription of contrasting intonational pitch accents (or, for that matter, lexical tones), it seems much less obvious that a system of "universal phonetic categories," ripe for precise transcription, is waiting to be discovered. Even the contrast principle is of less use here, insofar as we struggle in prosody to produce convincing analogues of the lexical minimal pairs that make contrast judgments possible

in segmental phonology. What kinds of things an International Prosodic Alphabet should aspire to encode, and indeed, whether such an alphabet is worth creating to begin with, are questions that at the moment of writing, elicit impassioned disagreement from otherwise even-tempered and well-adjusted researchers.

The question of how these phonetic categories in any case might connect to the contrastive phonological categories of the language has profound implications for a transcription system. An example of this issue can be found in the history of the development of the ToBI system, which took place over a series of four meetings in the early 1990s, first convened by Victor Zue of MIT. The intention was to develop a consensus transcription system for prosody, drawing on two sets of recently available resources: the emerging theory of the prosodic hierarchy, to be integrated with the emerging theory of intonation, and the accumulation of large databases of relatively spontaneous speech. These databases required hand-labeling of prosodic structure to enable machine learning of the signal characteristics associated with that structure, and there were a number of obstacles to the practical attainment of this goal. First, hand labeling is shockingly slow and costly, making it advantageous to pool labeled data, but second, different laboratories could not easily share their labeled data to form larger resources. That was because, over the previous half-century, different laboratories had developed idiosyncratic prosodic transcription systems, which did not allow for convenient sharing of labeled data. The goal of the workshops was to address this problem by developing a consensus system for prosodic transcription, so that laboriously labeled data could be more easily shared across laboratories and platforms.

During these developmental meetings, with a roomful of twenty to thirty meeting participants, all actively working on prosody or intonation but from widely different perspectives and with strikingly different goals, the discussion was lively, to say the least. One particularly striking division emerged between those who felt that the proposed system should capture all and only the prosodic aspects that were discernible in the acoustic signal of a given utterance, and those who felt that it should reflect the phonological contrasts in the prosody of the language, whether or not they were directly measurable in the signal, thereby enabling later analysis of the range of ways in which a given category can be signaled. In the absence of certainty about what the contrastive categories of English prosody are, those in the first camp were loath to label anything beyond what they saw in the signal, while those in the second camp were concerned that a purely surface approach to labeling would miss the point of the enterprise.

At times during the discussion, it seemed that only a thin veneer of civilized convention prevented the spilling of blood on the floor. The need for some kind of consensus was strong, however, and the outcome of these meetings, the ToBI transcription system, was a compromise. Some predictable aspects of the then-current phonology were eliminated in favor of capturing surface variation explicitly (e.g., explicit marking of downstep with the symbol !, rather than assuming downstepping of an H* pitch accent after a preceding bitonal accent), and some surface aspects of the acoustic signal were eliminated from explicit marking (e.g., declination across an intonational phrase, and relative scaling of accent-related F0 between phrases). The tension about the extent to which phonological analysis was baked into the system still remains problematic for many users and has led to proposals for alternative transcription systems that hew more closely to the surface form of the signal (e.g., RaP, and the ongoing effort to develop a phonetic-level IPrA).

As an aside, we note that a potential solution to this problem is to view prosodic transcription as a matter of specifying the set of individual acoustic-phonetic cues[13] to

the prosodic structure of an utterance. For example, a number of different cues to the presence of a prosodic boundary have been observed, including final lengthening, F0 excursion (edge tone), silent pause, F0 reset (up or down), and glottalization (initial or final in the word), and some preliminary reports suggest that these individual cues are processed by listeners. Brugos et al. (2018) have reported that the greater the number of cues present at a boundary, as identified by human labelers annotating individual cues, the more likely an independent set of ToBI labelers are to mark an intonational phrase boundary at that location. Along the same lines, Brugos et al. (2019) proposed a system for labeling individual correlates of prosodic disfluencies, based on work by Arbisi-Kelm (2006).

For purposes of symmetry with our presentation of the preceding questions, we should note that the question of prosodic transcription has not been of central concern to practitioners of the Fujisaki model. The model indeed is relevant to questions of transcription only in the sense that it makes assumptions about the kinds of things that are relevant phonologically in prosodic systems, and hence conceivably suggests that certain elements require transcription norms. Fujisaki and colleagues, to our knowledge, takes no official position on how best to transcribe prosody, and even suggests at one point (Mixdorff and Fujisaki 2000) that something like ToBI could serve as a system of category labels that would be interpreted phonetically in terms of the Fujisaki model.

As must be obvious by now, the purpose of the six questions we posed to our chapter authors was to reveal the goals of each of the major approaches to prosody that have been proposed over the past few decades. We hope that by highlighting the similarities and differences in the aims of each theory, the individual chapters will make it easier for practitioners to determine which approach best suits their needs. We leave it to our readers to decide the extent to which each of the following chapters achieves the level of explicitness that we believe it behooves us all to aspire to.

Notes

1. This is not to suggest the literal transfer of a single set of invariant motor or perceptual patterns from one word to another. Rather, each "reuse" of a category would be modulated by the full range of systematic, complex, contextual variability that has been documented in natural speech. The point is that all these patterns of variability, however expansive and multifaceted, would connect ultimately to the categories, as uttered in a particular pragmatic, prosodic, or other context rather than directly to a particular lexical item or semantic element. It is this intermediate connection that makes the problem tractable for language users.

2. This example, incidentally, also represents what we hold to be the pinnacle of the prosodist's refined art of composing lengthy carrier sentences out of nothing but sonorant segments. We salute its creator accordingly.

3. In the Fujisaki model, superposition, like everything else, is a matter of phonetic implementation. In autosegmental phonology, the process of tier conflation (McCarthy 1986) was hypothesized to be carried out during the phonological derivation and, potentially, could be extrinsically ordered relative to other phonological processes.

4. There are in fact certain tone strings that are often treated, informally at least, as though it were the entire contour with which utterance-level meanings were associated, making the contour more like the morpheme, and the tones that comprise it then more like phonemes—for example, the "contradiction contour," the "surprise-redundancy contour" (Sag and Liberman 1975), and the "Eastern European question contour" (Ladd [1996] 2008).

To the extent that the meanings of these contours are not built compositionally from their constituent tones, they would have a status something like that of phrasal idioms in syntax and semantics.

5. Actually, one persistent critique of the AM model (e.g., Mixdorff 2002) is that it does not in fact have an agreed-on, explicitly stated encoding function to begin with, and thus is difficult to evaluate in these terms. In any case, the example we are considering here is an entirely hypothetical one.

6. In fairness, it is also possible that most phonologists at the time had simply never heard of diphone synthesis, and, indeed, various proposals reminiscent of di- or triphones have in fact surfaced in the phonological literature (e.g., the tonal transitions of Dilley [2005] or the quantized contours of Q theory [Shih and Inkelas 2019]). Arguments for those proposals, however, were largely devoid of references to synthesis.

7. See, e.g., Jakobson's 1942 criticism of the "naive psychologism" of Baudouin de Courtenay and Saussure (Jakobson 1990, 224–230).

8. See the historic Principle 1 of the International Phonetic Association, first promulgated in 1888: "There should be a separate letter for each distinctive sound; that is, for each sound which, being used instead of another, in the same language, can change the meaning of a word" (Abramson 1988).

9. "It was quite clear to everybody at the Kiel convention that this is not the best possible chart; but it was also clear that we could not agree on how it should be improved" (Ladefoged 1990, on the discussions leading to the 1989 Kiel revisions to the IPA).

10. Catford (1990) played a central role in the enrichment of this portion of the vowel chart, which was officially carried out in 1993. Interestingly, at that time, one of these symbols was accidentally adopted in mirror image, owing to a typographical error. In 1996, this mistake was corrected, with the erroneous "closed epsilon," [ɘ], banished in favor of the intended "closed reversed epsilon," [ɞ]. Catford himself apparently preferred "barred open O" for this purpose, but was overruled for reasons that are not known to us (Pullum and Ladusaw 1996, 1998).

11. This is perhaps somewhat uncharitable. The substance of Catford's argument in favor of the proliferation of mid-central vowel qualities relied on a principle of parallelism within the system of Cardinal Vowels. If there are four total heights and three "horizontal" locations for vowel targets in this system, and with rounding distinctions separating primary and secondary versions of each, then we need at least four symbols to represent central vowels at heights two and three (open-mid and close-mid), in both round and unround variants. As a practical argument, however, he cites the need to distinguish "variants of the vowel in *bird"* at least at open-mid, mid, and close-mid levels. Note that any need to distinguish these things symbolically in transcription is not based on principles of contrast.

12. In some cases, what seems to happen is that transcribers fail to find a symbol that adequately reflects the sound in question, but they worry that choosing the closest option implies a level of precision and nuance that is unwarranted, leading them to opt out of IPA altogether, as when phonologists render Korean "tense" stops with an asterisk [such as p*], effectively meaning "this sound, in this language, whatever it actually is."

13. We use the term *cues* informally here, recognizing that some correlates of linguistic categories that are present in the signal may not be actively used by listeners or explicitly planned by speakers. Determining which correlates actually function in this way requires extensive experimentation, only some of which has been carried out to date.

References

Abramson, Arthur. 1988. "The Principles on Which the IPA Should Be Based." *Journal of the International Phonetic Association* 18 (2): 66–68.

Arbisi-Kelm, T. 2006. "An Intonational Analysis of Disfluency Patterns in Stuttering." PhD diss., UCLA.

Barnes, J. 2006. *Strength and Weakness at the Interface: Positional Neutralization in Phonetics and Phonology*. Berlin: Mouton de Gruyter.

Bell, C. C., H. Fujisaki, J. M. Heinz, and K. N. Stevens. 1961. "Reduction of Speech Spectra by Analysis-by-Synthesis Techniques." *Journal of the Acoustical Society of America* 33 (12): 1725–1736.

Brugos, A., M. Breen, N. Veilleux, J. Barnes, and S. Shattuck-Hufnagel. 2018. "Cue-Based Annotation and Analysis of Prosodic Boundary Events." In *Proceedings of the 9th International Conference on Speech Prosody*, June 13–16, 2018, Poznań, Poland, edited by Katarzyna Klessa, Jolanta Bachan, Agnieszka Wagner, Maciej Karpiński, and Daniel Śledziński, 245–249.

Brugos, A., A. Langston, S. Shattuck-Hufnagel, and N. Veilleux. 2019. "A Cue-Based Approach to Prosodic Disfluency Annotation." In *Proceedings of the 19th International Congress of Phonetic Sciences,* Melbourne, Australia, edited by Sasha Calhoun, Paola Escudero, Marija Tabain, and Paul Warren, 3413–3417.

Catford, J. C. 1990. "A Proposal Concerning Central Vowels." *Journal of the International Phonetic Association* 20:26–28.

Cheney, D. L., and R. M. Seyfarth. 1990. *How Monkeys See the World: Inside the Mind of Another Species*. Chicago: University of Chicago Press.

Chomsky, Noam, and Morris Halle. 1968. *The Sound Pattern of English*. New York: Harper.

Clements, G. N. 1991. "Place of Articulation in Consonants and Vowels: A Unified Theory." *Working Papers of the Cornell Phonetics Laboratory* 5:77–123.

Dilley, Laura C. 2005. "The Phonetics and Phonology of Tonal Systems." PhD diss., MIT.

Fant, Gunnar. 1960. *Acoustic Theory of Speech Production*. The Hague, Netherlands: Mouton.

Fujisaki, H. 1981. "Dynamic Characteristics of Voice Fundamental Frequency in Speech and Singing: Acoustical Analysis and Physiological Interpretations." In *Proceedings of the Fourth F.A.S.E. Symposium on Acoustics and Speech*, 2:57–70.

Fujisaki, H. 1988. "A Note on the Physiological and Physical Basis for the Phrase and Accent Components in the Voice Fundamental Frequency Contour." In *Vocal Fold Physiology: Voice Production, Mechanisms and Functions: Proceedings of the Fourth F.A.S.E. Symposium on Acoustics and Speech55*, edited by O. Fujimura, 347–355. New York: Raven.

Fujisaki, H. 1997. "Prosody, Models, and Spontaneous Speech." In *Computing Prosody*, edited by Y. Sagisaka, N. Campbell, and N. Higuchi, 27–42. Berlin: Springer-Verlag.

Fujisaki, H. 2004. "Information, Prosody, and Modeling—With Emphasis on Tonal Features of Speech." In *Speech Prosody 2004*, International Conference, Nara, Japan, March 23–26, 2004, edited by Bernard Bel and Isabelle Marlien, 1–10.

Fujisaki, H., and K. Hirose. 1982. "Modelling the Dynamic Characteristics of Voice Fundamental Frequency with Applications to Analysis and Synthesis of Intonation." *Preprints of Papers from the Working Group on Intonation, Thirteenth International Congress Linguists*, Tokyo, Japan, edited by Hiroya Fujisaki and Eva Gårding, 57–70.

Fujisaki, H., and K. Hirose. 1984. "Analysis of Voice Fundamental Frequency Contours for Declarative Sentences of Japanese." *Journal of the Acoustical Society of Japan (E)* 5 (4): 233–242.

Fujisaki, H., and S. Nagashima. 1969. "A Model for the Synthesis of Pitch Contours of Connected Speech." *Annual Report of the Engineering Research Institute, University of Tokyo* 28:53–60.

Fujisaki, H., S. Ohno, S., and C. Wang. 1998. "A Command–Response Model for F0 Contour Generation in Multilingual Speech Synthesis." In *Proceedings of the Third ESCA/COCOSDA International Workshop on Speech Synthesis*, 299–304.

Fujisaki, H., C. Wang, S. Ohno, and W. Gu. 2005. "Analysis and Synthesis of Fundamental Frequency Contours of Standard Chinese Using the Command-Response Model." *Speech Communication* 47:59–70.

Gårding, E. 1983. "A Generative Model of Intonation." In *Prosody: Models and Measurements*, edited by A. Cutler and D. R. Ladd, 11–25. Berlin: Springer.

Glewwe, E., J. Zymet, J. Adams, R. Jacobson, A. Yates, A. Zeng, and R. Daland. 2018. "Substantive Bias and the Acquisition of Final (De)Voicing Patterns." Paper presented at the 92nd Annual Meeting of the Linguistic Society of America, Salt Lake City, Utah, January 4–7.

Gordon, Matthew. 2016. *Phonological Typology*. Oxford: Oxford University Press.

Grice, Martine, D. Robert Ladd, and Amalia Arvaniti. 2000. "On the Place of 'Phrase Accents' in Intonational Phonology." *Phonology* 17:143–185.

Halle, Morris, and Kenneth Stevens. 1959. "Analysis by Synthesis." In *Proceedings of the Seminar on Speech Compression and Processing*, vol. 2, paper D-7, edited by W. Wathen-Dunn and L. E. Woods.

Hirst, D. 2011. "The Analysis by Synthesis of Speech Melody: From Data to Models." *Journal of Speech Sciences* 1 (1): 55–83.

Hockett, Charles F. 1958. *A Course in Modern Linguistics*. New York: Macmillan.

Hyman, Larry M. 2018a. "What Is Phonological Typology?" In *Phonological Typology*, edited by Larry M. Hyman and Frans Plank, 1–20. Berlin: Mouton de Gruyter.

Hyman, Larry M. 2018b. "Why Underlying Representations?" *Journal of Linguistics* 54 (3): 591–610.

Jakobson, Roman. 1990. "The Concept of the Phoneme." In *On Language*, edited by Linda R. Waugh and Monique Monville-Burston, 217–242. Cambridge, MA: Harvard University Press.

Kiparsky, Paul. 2006. "The Amphichronic Program vs. Evolutionary Phonology." *Theoretical Linguistics* 32 (2): 217–236.

Ladd, D. Robert. (1996) 2008. *Intonational Phonology*. 2nd ed. Cambridge: Cambridge University Press.

Ladd, D. Robert. 2014. *Simultaneous Structure in Phonology*. Oxford: Oxford University Press.

Ladefoged, Peter. 1990. "The Revised International Phonetic Alphabet." *Language* 66 (3): 550–552.

Ladefoged, Peter. 2005. "Features and Parameters for Different Purposes." *UCLA Working Papers in Phonetics* 115 (104): 1–13.

Liu, F, Y. Xu, S. Prom-on, and A. C. L. Yu. 2013. "Morpheme-like Prosodic Functions: Evidence from Acoustic Analysis and Computational Modeling." *Journal of Speech Sciences* 3:85–140.

Martinet, André. 1949. "La double articulation linguistique." *Travaux du Cercle Linguistique de Copenhague* 5:30–37.

McCarthy, John. 1986. "OCP Effects: Gemination and Antigemination." *Linguistic Inquiry* 17 (2): 207–263.

Mixdorff, Hansjörg. 1998. "Intonation Patterns of German—Model-based Quantitative Analysis and Synthesis of f0-Contours." PhD diss., Technische Universität Dresden.

Mixdorff, Hansjörg. 2002. "Speech Technology, ToBI and Making Sense of Prosody." In *Speech Prosody 2002, International Conference*, Aix-en-Provence, France, April 11–13, 2002; ISCA Archive. http://www.isca-speech.org/archive/sp2002.

Mixdorff, Hansjörg, and Hiroya Fujisaki. 2000. "Symbolic versus Quantitative Descriptions of F0 Contours in German: Quantitative Modelling Can Provide Both." In *Prosody 2000: Speech Recognition and Synthesis*, Krakow, Poland, October 2–5.

Moreton, Elliott. 2008. "Analytic Bias and Phonological Typology." *Phonology* 25 (1): 83–127.

Moreton, Elliott, and Joe Pater. 2012. "Structure and Substance in Artificial-Phonology Learning, Part II: Substance." *Language and Linguistics Compass* 6 (11): 702–718.

Öhman, S., and J. Lindqvist. 1965. "Analysis-by-Synthesis of Prosodic Pitch Contours." *Speech Transmission Laboratory: Quarterly Progress and Status Reports* 6 (4): 1–6.

Pierrehumbert, J. 1980. "The Phonology and Phonetics of English Intonation." PhD diss., MIT.

Pierrehumbert, Janet, and Julia Hirschberg. 1990. "The Meaning of Intonational Contours in the Interpretation of Discourse." In *Intentions in Communication*, edited by P. R. Cohen, J. Morgan, and M. E. Pollack, 271–311. Cambridge, MA: MIT Press.

Prince, Alan, and Paul Smolensky. 1993. *Optimality Theory: Constraint Interaction in Generative Grammar*. Technical Report 2. New Brunswick, NJ: Rutgers University Center for Cognitive Science.

Pullum, Geoffrey K., and William A. Ladusaw. 1996. *Phonetic Symbol Guide*. Chicago: University of Chicago Press.

Pullum, Geoffrey K., and William A. Ladusaw. 1998. "Vowel Charts and Central Vowel Transcriptions in American and IPA Traditions." *Publication of the American Dialect Society* 80 (1): 5–33.

Sag, Ivan, and Mark Liberman. 1975. "The Intonational Disambiguation of Indirect Speech Acts." In *Papers from the Eleventh Regional Meeting, Chicago Linguistic Society*, edited by Robin E. Grossman, L. James San, and Timothy J. Vance, 487–497.

Sandler, W., M. Aronoff, I. Meir, and C. Padden. 2011. "The Gradual Emergence of Phonological Form in a New Language." *Natural Language and Linguistic Theory* 29:503–543.

Sapir, E. 1933. "The Psychological Reality of Phonemes." In *Selected Writings of Edward Sapir in Language, Culture and Personality*, edited by D. G. Mandelbaum, 46–60. Berkeley: University of California Press.

Shih, Stephanie, and Sharon Inkelas. 2019. "Autosegmental Aims in Surface-Optimizing Phonology." *Linguistic Inquiry* 50 (1): 137–196.

Ternes, E., and T. Vladimirova-Buhtz. 1990. "Bulgarian." *Journal of the International Phonetic Association* 20 (1): 45–47.

Yip, M. J. 1980. "The Tonal Phonology of Chinese." PhD diss., MIT.

The MIT Press would like to thank the anonymous peer reviewers who provided comments on drafts of this book. The generous work of academic experts is essential for establishing the authority and quality of our publications. We acknowledge with gratitude the contributions of these otherwise uncredited readers.

This book was set in Stone Serif and Stone Sans by Westchester Publishing Services.

Library of Congress Cataloging-in-Publication Data is available.
Names: Barnes, Jonathan, 1970– editor. | Shattuck-Hufnagel, Stefanie, editor.
Title: Prosodic theory and practice / edited by Jonathan Barnes and Stefanie Shattuck-Hufnagel.
Description: Cambridge, Massachusetts : The MIT Press, 2022. | Includes bibliographical references and index.
Identifiers: LCCN 2021000764 | ISBN 9780262543170 (paperback)
Subjects: LCSH: Prosodic analysis (Linguistics)
Classification: LCC P224 .P739 2022 | DDC 414/.6—dc23
LC record available at https://lccn.loc.gov/2021000764