Book Reviews

Prosody: Theory and Experiment. Studies presented to Gösta Bruce

Merle Horne (editor)
(University of Lund)

Dordrecht: Kluwer Academic Publishers (Text, speech and language technology series, edited by Nancy Ide and Jean Véronis, volume 14), 2000, v+358 pp; hardbound, ISBN 0-7923-6579-8, $153.00, £95.00, Dfl 290.00

Reviewed by
Chilin Shih, Bell Labs
and
Richard Sproat, AT&T Labs – Research

This volume consists of a collection of solicited articles, presented as a tribute to Gösta Bruce, one of the leading researchers in the field of prosody. The papers are for the most part review articles (as perhaps befits a volume of this kind), and the volume thus serves an important function in prosody research in giving a broad snapshot of the field as it stands at the dawn of the new millennium. The coverage of the volume is indeed broad, including such areas as intonation inventory (Pierrehumbert, Ladd, Hirst et al., Gussenhoven), tune-text alignment (Pierrehumbert, Ladd, Gussenhoven), acoustic correlates of accent and stress (Terken et al., Beckman et al.), prominence (Terken et al., Ladd), prosodic structure and juncture (Hirst et al., Shattuck-Hufnagel, Selkirk, Ostendorf), timing (Beckman et al., Campbell), and differences between speaking styles (Hirschberg). The book is successful in providing readers with a reasonably clear perspective on what the relevant questions are, what kind of evidence is available to answer those questions, and who has done what. The authors hold in common the view that to understand prosody one needs to investigate abstract representations: one studies intonational categories and prosodic structures, rather than, for instance, vectors of f0 values.

The book starts with Horne's introduction, which introduces the articles, and puts them in perspective, especially with respect to the work of Gösta Bruce.

Pierrehumbert's article is an excellent survey of her own model of intonation, and would be very suitable for an introductory course on intonation: in forty pages one gets a condensed version of the historical development of a major intonational framework. It outlines the major issues that have changed the views of intonation researchers from the structuralists to today, with a section describing Gösta Bruce's influential work on Swedish intonation, as well as his contribution to the development of intonation theory embodied in the concept that sentence intonation can be represented as a sequence of tones.

In Pierrehumbert's intonation system, one of the central theoretical issues is how to interpret continuous phonetic data and link it to discrete phonological categories. She reviews arguments for the two-level representation of tones, using only H (high) and

L (low), as well as the arguments surrounding the assignment of accent categories. Downstep data provides a crucial argument against multiple levels of tonal representation. The argument goes as follows: Each successive downstepped tone clearly sounds distinct, and in some languages the distinctions are phonological. It can also be demonstrated that one can have as many downstepped pitch levels as are allowed by performance factors such as utterance length. If one were to represent each distinct level phonologically, then the conventional four levels proposed early on by structuralists, or any predetermined number of levels for that matter, are not enough. A more coherent treatment is to use minimal phonological levels—in this case, H and L—to implement the phonemic distinctions and phonetic models to predict gradient pitch height. The downstep data has been successfully treated in this way.

The same philosophy applies to the classification of accent inventories. Nuclear accent and pre-nuclear accent in English declarative sentences have different peak alignment patterns. Is it possible that they belong to the same underlying accent category? Pierrehumbert suggests that the answer is yes and reviews arguments supporting this treatment. Many factors affect the alignment pattern of the f0 peak with the stressed syllable. But there is no need to posit a different category whenever one sees surface variations, so long as the surface variations can be predicted by a phonetic model from a unique phonological representation.

It is interesting to read Ladd's paper right after Pierrehumbert's. Ladd plays devil's advocate and raises three questions directly addressing the issue of intonation representation in Pierrehumbert's framework:

1. What is a tone? There is a level of abstraction in Pierrehumbert's tonal inventory involving mismatches in the assignment of H and L on the one hand, and turning points in surface pitch contours on the other hand. Should one take the turning points in surface intonation contour more seriously?
2. What is the meaning of the starred tone, as in the "*" of H*+L or L*+H? Ladd suggests that the starred tone should be the landing site of emphasis, and he expects H* to go higher and L* to go lower under emphasis. Under this assumption, he raises the issue of Bruce's (1987) re-analysis of his Swedish data, which assigned H+L* to accent I and H*+L to accent II. Fant and Krukenberg (1994) reported that, under focus, the pitch of the low tone in accent II (H*+L) is lower, while the pitch of all targets in accent I (H+L*) are raised.
3. Is the tune-text association convention as mediated by the star really meaningful? Ladd raises the question of Greek L+H tone, where L is aligned right before the stressed syllable and H is aligned right after the stressed syllable. Ladd's concern is that, in this case, nothing is actually aligned with the stressed syllable. It is also implied that in languages such as English or Greek, the tune-text association may not be as rigid as in a tone language such as Yoruba (see below).

Ladd's questions reiterate the theme raised by Pierrehumbert on the difficulty of separating phonetic variation from phonological representations. In a sense, the Greek L+H case remains true to the autosegmental spirit of tune-text association: the star of the text provides an anchor, and alignment of the tune is made with reference to this anchor. As long as the facts are clear, an alignment model can be built successfully, referring to the whole contour or to any number of points along the contour. The

alignment model can be quite complicated (see van Santen and Möbius [2000]) and the predicted tone landing site can be quite far from the anchor.

In addition, it turns out that the tune-text misalignment in a tone language can be substantially more dramatic than the Greek case described by Ladd. In Mandarin Chinese (Shih and Kochanski 2000; Xu 2001), it is not uncommon to have a tone target shifted completely off the syllable it originates from. In Yoruba (Laniran 1992), a target can be delayed by several syllables. In a tone language, both the tonal inventory and the ideal, phonological tune-text alignment are known. So when the tune and text alignment are off by a few syllables, one has no choice but to acknowledge the misalignment and to zero in on an alignment model predicting the alignment pattern. When dealing with a non-tone language, it is not as easy to ascertain a case of long-distance misalignment and one often ends up with a phonological analysis that is closer to surface observables.

The paper of Hirst, Di Cristo, and Espesser is an overview of their intonation model, providing a contrast to the ToBI-based articles in the book. Theirs is an automated intonation model used for speech analysis and synthesis. The MOMEL algorithm analyzes f0 curves, smoothing out some of the micro-prosody and finding target points in the f0 contours. The target points are converted to the INTSINT transcription system. Basically, INTSINT defines the highest and lowest targets in the utterance as H (high) and B (bottom), respectively. Other phrase-initial targets are labeled as M (mid). The rest of the targets in the utterance are assigned labels reflecting relative relations with preceding and following targets: L (low), U (up), D (down), S (same). It should be apparent from this description that the analysis part of the system is language-independent: it depends on speech signals rather than on language knowledge. For the purpose of synthesis, some level of language knowledge is needed to write or to train a language-dependent intonation grammar. The fO contours will then be generated from the targets predicted by the grammar.

Terken and Hermes provide a comprehensive review of the literature on prominence perception; readers get a clear sense of what questions should be asked and have been asked, even though the answers may not always be clear. One central issue addressed in this article is the question of what makes two accents sound equally prominent. Interesting experiments have been done comparing different pitch accents, as well as pitch accents in different pitch registers (e.g., male vs. female), in different positions of a sentence, and with or without a declining baseline. Technical questions revolve around what the correct thing is to measure (pitch excursion, or pitch-level difference), and in which scale (Hz, BART, Mel, or ERB). The data support a general declination model, in which early accents are bigger and higher than later, equally prominent accents. The tilt of the baseline has an effect on the perception of prominence. Among different pitch accents with equal excursion size, falling pitch accents lend more prominence than rising or rise-fall accents.

Gussenhoven's contribution presents new work on a lexical tone contrast in the Roermond dialect of Dutch. Roermond Dutch, like other Limburg dialects, has a lexical accentual system reminiscent of that of Swedish. There are two types of accent. Words with Accent I have no lexically prespecified tone. Accent II words, following Gussenhoven's analysis, have a high (H) tone linked to the second mora of the accented syllable: note, therefore, that Accent II can only occur in words where the syllable that would bear the accent is bimoraic. Accent II displays a number of interesting features. First, if no boundary tone and no pitch accent is associated with the accented syllable, the H does not surface. Second, if a low (L) pitch accent is associated with the syllable, the H transmutes to a L so that you get a sequence L*L. Third, and most interesting, a boundary tone—L_i or H_iL_i—apparently shows up *before* the Accent II high. So an

"underlying" sequence such as H*HL$_i$—where H* is an intonationally assigned pitch accent, H is the Accent II tone, and L$_i$ is the boundary tone—surfaces as an HLH tone sequence rather than the expected HHL. As Gussenhoven notes, this is the first documented instance of a boundary tone being anything other than peripheral, and it has somewhat the flavor of morphological infixation.

Gussenhoven argues that a derivational account—what he terms an "SPE" (Chomsky and Halle 1968) account—would lead to an ordering paradox between two rules, both of which one would seemingly need: the rule that transmutes L*H to L*L, and the metathesis rule that reorders the boundary tone before an Accent II H. Note that in this derivational account, Gussenhoven assumes that the metathesis rule is a transformational rule that refers to the tone sequence and the boundary, and nothing else. Having rejected such an account, Gussenhoven presents a constraint-based analysis within the Optimality Theory (OT) framework, making use of about ten ranked constraints. Central to the boundary tone reordering is the assumption of two constraints, one (ALIGNT$_i$RT) that states that the boundary tone wants to align to the right of its phrasal domain, and the other (ALIGNLEXRT) which states that the Accent II H wants to align to the right of its syllable, which of course coincides with the right of the phrasal domain if that syllable is final. The ranking ALIGNLEXRT $\gg$ ALIGNT$_i$RT achieves the desired result that the Accent II H comes after the boundary tone.

Though Gussenhoven presents what seems to be a particularly compelling argument for OT, the Roermond data are actually grist for any number of theoretical mills. For example, taking Gussenhoven's proposal for a lexical H tone at face value, one could explore the possibility of a more traditional autosegmental analysis: the straw man derivational analysis that Gussenhoven presents is hardly fair to the quarter of a century of phonology between SPE and the advent of OT. Then there is the possibility of *not* taking Gussenhoven's analysis at face value. Indeed, two properties of Accent II—the transmutation of the H to L after L*, and the complete loss of H in accent-free non-boundary contrasts—suggests the possibility that Accent II may not involve a lexically specified H at all, but rather merely a different timing specification for whatever accent (if any) gets associated with the syllable. Such an approach would not be without complications, but it seems nonetheless worth exploring.

Beckman and Cohen's is the second of the articles in this book to report new data. This article is a follow up of earlier work by Beckman and Edwards (1994) on the differences in the jaw-opening movements of two types of lengthening: lengthening for accent, and phrase-final lengthening. The data consist of the syllable *pop*, stressed/accented, stressed/unaccented, and unstressed. These stimuli are embedded in phrase-final as well as non-final positions. Beckman and Cohen consider three articulatory models to account for the jaw-tracing differences between full and reduced vowels: a truncation model, a rescaling model, and a hybrid model. The preliminary analysis suggests that the hybrid model works best. This study supports earlier findings that not all lengthenings are accomplished in the same way. Contrasting with an unstressed syllable, a stressed syllable is longer, and has a more extreme displacement of the jaw, but with higher velocity in the movement; this is not accounted for by the truncation model, which predicts the velocity of the movement to be the same. In phrase-final position, lengthening is accompanied by slower movement.

Shattuck-Hufnagel's paper presents three types of arguments, based on stress shift, glottalization, and rhythmic pattern in speech—with data obtained from a speech corpus—to support the **prosodic planning hypothesis**, namely that speech production is planned with reference to prosodic structure. She suggests that the so-called stress-shift rule is not really a rule that shifts stress to avoid stress clash. The main argument is that the "stress shift" effect may be achieved by the addition of an accent without

shifting the original stress; it may also occur when there is no stress clash. There is a tendency for speakers to use a pair of accents to frame a prosodic phrase. The first landing site is the earliest full vowel in the phrase, and the last one is the nuclear accent. The stress-shift rule can be subsumed under this mechanism. Glottalization of a vowel-initial word such as *apple* is more common in phrase-initial position. Also, phrase-final position is frequently marked by glottalization. The use of accents and glottalization both have the effect of framing a prosodic phrase. Further evidence for prosodic structure comes from the preference for the use of alternating stress. Although sentences in natural speech often do not show strictly alternating stress, comparisons of sentences with and without rhythmic patterns show that sentences with rhythmic patterns are easier to produce and less prone to speech errors.

Selkirk presents an interesting analysis of English phonological phrasing in terms of OT. She starts by reviewing evidence from Bantu languages for the universality of two constraints, namely ALIGN_{R} XP, which states that "the right edge of any XP (maximal projection) in syntactic structure must be aligned with the right edge of a MaP (major phrase) in prosodic structure", and a constraint proposed by Truckenbrodt (1995), WRAP XP, which states that "the elements of an input morphosyntactic constituent of type XP must be contained within a prosodic constituent of type MaP in output representation". Since the first constraint requires the right edges of MaPs to align with the right edge of each XP, whereas the second requires all XPs to be contained within a (single) MaP, the two constraints are inherently in conflict. In English, it seems that there is no evidence for ranking between ALIGN_{R} XP and WRAP XP. This is because a sentence such as *She lóaned her róllerblades to Róbin*—where, crucially, each of the words *loaned*, *rollerblades* and *Robin* are accented—can be phrased as either *(She lóaned her róllerblades)*$_{\text{MaP}}$ *(to Róbin)*$_{\text{MaP}}$, with two MaPs, or *(She lóaned her róllerblades to Róbin)*$_{\text{MaP}}$, with one. The first case violates WRAP XP, the second violates ALIGN_{R} XP, but both violations seem to be equal. What is not accounted for is the failure of an "overphrased" version, namely *(She lóaned)*$_{\text{MaP}}$ *(her róllerblades)*$_{\text{MaP}}$ *(to Róbin)*$_{\text{MaP}}$, which is a violation of WRAP XP, but not ALIGN_{R} XP. This motivates the introduction of a third, lower-ranked constraint BINMAP, which requires MaPs with just two accentual phrases; the overphrased version then has three violations of this constraint.

Selkirk then turns to ALIGN_{R} FOCUS, which requires the alignment of a MaP with a focused constituent, and she argues that it is higher ranked than the other constraints. Thus a focused version of *She* ***lóaned*** *her róllerblades to Róbin* has one optimal candidate: *(She* ***lóaned)***$_{\text{MaP}}$ *(her róllerblades)*$_{\text{MaP}}$ *(to Róbin)*$_{\text{MaP}}$. Here, the boundary after *loaned* is favored by ALIGN_{R} FOCUS, and the boundary after *rollerblades* is favored by ALIGN_{R} XP.

Selkirk's data are based on intuitive judgments concerning the putative presence or absence of a MaP boundary tone, the diagnostic she adopts (following Beckman and Pierrehumbert [1986]) for deciding whether a phrase boundary is present. It would be interesting to see to what extent these data hold up under experimental conditions.

Ostendorf presents a brief review of linguistic and engineering issues related to the automatic detection of prosodic boundaries. As she points out, automatic boundary detection is desirable in a number of areas of speech technology. For example, in speech recognition and understanding, prosodic information can, in principle, be used to prune the search space (since some hypotheses are likely to be incompatible with a given phrasing) and to score linguistic hypotheses. Automatic detection is also desirable in text-to-speech synthesis to aid in the rapid development of prosodically labeled databases. A holy grail of this enterprise, as Ostendorf notes, is an approach that is robust enough to work on spontaneous speech: current automatic phrase de-

tection methods work well only on read speech and perform considerably less well on the kind of conversational speech found in the Switchboard corpus.

Campbell presents an overview of work on duration modeling, starting with a rather detailed account of Klatt's (1973) rule-based model, and covering various statistical approaches such as Riley's (1990) CART-based models and van Santen's (1994) sum-of-products models. Campbell's own view is that segment-based models of duration are misguided because they are based on the notion of a segment's "inherent duration," and that instead one should model higher levels of prosodic structure (syllables, feet, or even prosodic phrases), deriving segmental durations once the higher-level durations are set up. The second half of the article describes earlier work of Campbell that provides support for a syllable-based approach from English and Japanese.[1]

Hirschberg concludes the book with a review of earlier work on prosodic cues that differentiate speaking style—or, more properly, two particular speaking styles, namely read speech and spontaneous speech. She catalogs differences in rate (read speech is faster), differences in the distribution of different boundary tones, and differences in the rates of disfluencies, as well as a few other factors. Disfluencies, while more common in spontaneous speech, as one might expect, are nonetheless sufficiently rare in both styles that they are not a particularly useful cue to distinguishing the two.

Finally, a word on production quality, which unfortunately is mediocre. There are a variety of problems, particularly with the presentation of some of the figures and the equations. So, the shaded areas of the tableaux in Selkirk's paper are too dark, though the ones in Gussenhoven's paper are fine. In at least a couple of the papers—Gussenhoven, Campbell—there are some quite annoying changes in font size between successive linguistic examples or equations. On the whole the production quality is not what you would expect for a volume that lists at over US$150. But there is presumably nothing to do here but lament the fact that as academic publishers continue to up the prices of their wares, they also seem to be taking less and less care in their production.

References

Beckman, Mary and Jan Edwards. 1994. Articulatory evidence for differentiating stress categories. In P. A. Keating, editor, *Phonological Structure and Phonetic Form: Papers in Laboratory Phonology III*. Cambridge University Press, Cambridge, pages 7–33.

Beckman, Mary and Janet Pierrehumbert. 1986. Intonational structure in English and Japanese. *Phonology*, 3:255–309.

Bruce, Gösta. 1987. How floating is focal accent? In K. Gregersen and H. Basbøll, editors, *Nordic Prosody IV*. Odense University Press, Odense, pages 41–49.

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

Fant, Gunnar and A. Krukenberg. 1994. Notes on stress and word accent in Swedish. *Speech Transmission Laboratory Quarterly Status Report*. Technical report, Department of Speech Communication and Music Acoustics, KTH, Stockholm.

Klatt, Dennis. 1973. Interaction between two factors that influence vowel duration. *Journal of the Acoustical Society of America*, 54:1102–1104.

Laniran, Yetunde. 1992. *Intonation in Tone Languages: The Phonetic Implementation of Tones in Yoruba*. Ph.D. thesis, Cornell University, Ithaca, NY.

Riley, M. D. 1990. Tree-based modeling for speech synthesis. In *Proceedings of the ESCA Workshop on Speech Synthesis*, 229–232, European Speech Communication Association.

Shih, Chilin and Greg Kochanski. 2000. Chinese tone modeling with Stem-ML. In *International Conference on Spoken Language Processing, 2000*, article 1232, Beijing.

Truckenbrodt, H. 1995. *Phonological Phrases: Their Relation to Syntax, Prominence and Focus*. Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, MA.

1 In work published about the same time as this book, van Santen and Shih (2000) present arguments against the syllable-based approach as espoused by Campbell.

van Santen, Jan. 1994. Assignment of segmental duration in text-to-speech synthesis. *Computer Speech and Language*, 8:95–128.

van Santen, Jan and Bernd Möbius. 2000. A quantitative model of f0 generation and alignment. In A. Botinis, editor, *Intonation: Analysis, Modelling and Technology*. Kluwer Academic Publishers, Dordrecht, pages 269–288.

van Santen, Jan and Chilin Shih. 2000. Suprasegmental and segmental timing models in Mandarin Chinese and American English. *Journal of the Acoustical Society of America*, 107(2): 1012–1026.

Xu, Yi. 2001. Fundamental frequency peak delay in Mandarin. *Phonetica*, 58:26–52.

Chilin Shih is a Research Scientist at Bell Laboratories, Lucent Technologies. She has built text-to-speech systems for many languages. Her current research focuses on tone and intonation modeling. Shih's address is 600 Mountain Avenue, Murray Hill, NJ 07974; e-mail: cls@research.bell-labs.com. *Richard Sproat* is a Technology Consultant at AT&T Labs – Research. He has done research in computational morphology, writing systems, and text-analysis for text-to-speech synthesis, where he has worked on accent prediction and multilingual text analysis. Sproat's address is 180 Park Avenue, Florham Park, NJ 07922; e-mail: rws@research.att.com.