Modeling the Production of Prosody

Mary E. Beckman
Department of Linguistics, Ohio State University
222 Oxley Hall, 1712 Neil Ave., Columbus, OH 43210-1298, USA

ABSTRACT

This tutorial reviews some of the insights gained from fifteen years of research on the production of prosodic categories, concentrating primarily on modeling fundamental frequency patterns as reflections of intonation and phrasing. It closes with a hint of some at the directions in which this work is being extended to understanding the function of prosody in discourse and to the modeling of other phonetic dimensions.

INTRODUCTION

This tutorial was originally intended to introduce the "Production" of prosody. In the course of assigning papers to the different sections of this ESCA Workshop on Prosody, however, the organizers have expanded the scope of this section to cover "Models" and "Synthesis" as well. I can think of no more appropriate expansion. There is a particularly intimate and necessary connection between production data and modelling. Absent the kind of explicit laying out of theoretical assumptions afforded by a well-developed model, particularly by a model which has undergone that most rigorous test of being implemented in a synthesis system, production data are notoriously difficult to interpret. To be sure, this difficulty of interpretation is characteristic of all speech production data, but it is compounded for prosodic data, where there is not always the immediate theory of segmentation and contrast that the phonemic principle gives us. That is, for most aspects of prosodic structure, the contrasts are not the salient lexical contrasts that gave us the International Phonetic Alphabet for consonants and vowels. Rather, prosodic contrasts usually involve the pragmatic relationship of a syllable, word, or phrase to other elements in the utterance or discourse context, the sort of thing that is much harder to introspect about than the fact that one word is not another word. A speech scientist who wants to investigate some aspect of prosody must, therefore, first consciously decide to adopt some model to guide the design of the investigation and the interpretation of its results.

MODELING FUNDAMENTAL FREQUENCY PATTERNS

Nowhere are the connections between data and model building and between models and synthesis more fully exercised than in the investigation of phonologically contrastive intonational events as they are realized in fundamental frequency patterns. The location of this ESCA Workshop on Prosody in Lund is a felicitous reminder of the models of Swedish intonation developed here. The decade and a half since Bruce's (1977) seminal monograph has given us comprehensive descriptive models and systems for synthesizing fundamental frequency patterns in many other languages as well. Indeed for some of these languages, we have the luxury of two or more competing descriptions. For example, for English, we have Pierrehumbert's model implemented in the AT&T Bell Labs text to speech system (Anderson et al. 1984), Ladd's model implemented in the University of Edinburgh CSTR system (Ladd, 1987), and even an IPO-style system (Willems, Collier, and 't Hart, 1988). For Dutch, we have of course the IPO model itself ('t Hart and Collier, 1973), and a model implemented at Nijmegen (Gussenhoven and Rietveld, 1992). For Mandarin Chinese, there have been at least three different models (Shih, 1988; Gårding, 1987; Fujisaki, Hirose, Halle, and Lei, 1990), and likewise for Japanese (Fujisaki and Sudo, 1971; Sagisaka, 1990; Pierrehumbert & Beckman, 1988). The number of languages which have undergone the analysis and experiments necessary

to construct synthesis models increases yearly. The system for German by Kohler (1991) and for Italian by Avesani (1990) are only two of many recent examples.

A notable result of this decade and a half of work is that we can now say with a fair degree of confidence which aspects of fundamental frequency patterns are likely to generalize across languages, and which aspects are likely to vary. One of the more important generalizations is that fundamental frequency patterns in all languages seem to be related in some way to discourse organization. In general, coherence among words or phrases can be signaled when each following F0 peak is systematically reduced relative to preceding peaks (as opposed to disjuncture, signaled when a following peak is clearly not in such a systematic relationship — e.g. by being produced in an expanded pitch range). In addition, some languages encode aspects of this general tendency in the phonology, by imposing strictly phonological constraints on the F0 downtrend across coherent sequences of phrases. For example, in both standard (Tokyo) Japanese and English, a large component to the downtrend is phonologically conditioned: each occurrence of a sequence of tones that is grouped together functionally into a "pitch accent" triggers a particularly large "downstep" of following peaks relative to utterances which do not have a downstep-triggering pitch accent in the same position. Moreover, in both languages, the downstep seems to be limited in application to a phonological constituent that is intermediate in the prosodic hierarchy between the prosodic word and the intonational phrase and a "reset" of the pitch range occurs at phrasal boundaries (Pierrehumbert and Beckman, 1988). This similarity is striking when we consider the very different distribution and function of the pitch accent in the two languages. In Japanese, the presence of a pitch accent on a particular syllable is part of the dictionary specification (accented words contrast lexically with unaccented), whereas in English, pitch accents are pragmatic markers associated to the most stressed syllable in words that are particularly salient in the current discourse segment. Yet the two languages are alike in that every bitonal pitch accent triggers downstep.

However, this kind of downstep is not a language universal. In some other languages, the trigger is different. For example, in Mandarin Chinese, downstep occurs each time there is a syllable bearing one of the three lexical tones other than the high level tone (Shih, 1988). In still other languages, there is no identifiable phonologically triggered component of the downtrend. For example, in Standard Danish (Thorsen, 1980) and Chonnam Korean (Jun, 1989), something like downstep seems to be triggered by each pitch accent, but there are no miminal pairs of accent type or of accented versus unaccented words to distinguish a potential phonological trigger from the more general pragmatic considerations of signaling discourse coherence. Bruce (1982) has shown that a similar reduction of successive accent peaks occurs in southern Swedish as well, but unlike in Danish and Korean, Swedish "downstep" is further limited to the region of the phrase after the nuclear stress (see below). Thus, by careful phonological and phonetic investigation of the intonational systems of many languages, we come across a broad generalization — that many languages use a reduction of later F0 peaks relative to earlier ones within a constituent at some level of the prosodic hierarchy to signal discourse coherence — but that languages differ in whether or how phonological contrasts condition the downtrend.

PITCH ACCENTS AND STRESS

Another notable result of this decade and a half of work on modeling F0 patterns in many languages is that we can now build more useful taxonomies. In comparing Swedish and Japanese with English and French in these F0 models, we see that the old impressionistic dichotomy between "pitch accent languages" and "intonation languages" is not so compelling. Whereas the older taxonomy classified Swedish and Japanese together in opposition to English and French because both have lexically determined pitch contrasts (accent 1 words versus accent 2 words in Swedish, accented versus unaccented words in Japanese), we now might note that all four languages have pitch patterns that fit the definition of "pitch accent" — namely, a tone or closely connected sequence of tones that

is associated phonologically to some designated syllable within a word. Thus, they all are "pitch accent languages" in some sense. They all also are "intonation languages" in the sense of using tone patterns and pitch range relationships to group together words into prosodic phrases of various sizes: all four languages have tonally marked intonational phrases, and French and Japanese also have a smaller tonally delimited minor phrase (the French "rhythm group" and the Japanese "accentual phrase").

Given these similarities, the dimension of lexical contrast upon which the older taxonomy hinged now seems a grab bag of fundamentally unrelated characteristics. In Swedish, the lexical contrast between accent 1 and accent 2 is primarily one of pitch accent shape whereas in Japanese the lexical contrast involves pitch accent placement alone and shape is not distinctive. Both these types of lexical contrast now seem considerably less useful in categorizing the prosodic systems of these languages than the relationship of pitch accent to sentence rhythm. In both Swedish and English, a basic aspect of the rhythm of an utterance is the pattern of alternation between strong (stressed) and weak (unstressed) syllables, with pitch accent placement functioning to mark the strongest stressed syllables. They both contrast in this to French and Japanese, where the salient rhythm is instead a grouping of syllables into tonally marked minor phrases, with no very compelling relationship between pitch accent and syllable prominence.

Categorizing the languages on these lines leads, first of all, to a better understanding about the relationship between prosodic structure and segmental contrasts. Accented syllables (or accentable syllables) in Swedish and English are very different from unaccented syllables. The consonants in accented syllables are "stronger" (e.g., voiceless stops are strongly aspirated), vowels are longer and more fully realized (i.e., closer to the periphery of the vowel triangle), and, unlike the vowels in unstressed syllables, they cannot be reduced to the point of apparent deletion (cf. Fokes and Bond, 1993). De Jong (1991) characterizes these differences between accented and unaccented syllables in English as one of local "hyperarticulation": in languages such as English, accented syllables have special status in the conflict between the needs of the speaker to minimize effort and the needs of the hearer to maximize distinctive (Lindblom, 1990). In Japanese, by contrast, accented syllables are not different in length (see, e.g., Beckman, 1986) and seem hardly different in propensity to vowel "reduction". French is somewhat more akin to English and Swedish in that the (normally final) accented syllable is longer in duration than the syllables preceding it in each "rhythm group". However, closer inspection of the kinematic patterns involved shows that this lengthening is more like "pre-boundary" lengthening than it is like the accentual effect in English (Fletcher and Vatikiotis-Bateson, 1991). As Martin (this volume) puts it, the characteristics of "stress" in French are "particularly elusive" by comparison to those in English or Swedish.

Categorizing Swedish with English and French with Japanese on these grounds also leads to predictive insights about the ways in which such discourse categories as "broad versus narrow focus" will be realized in the prosody of naturally occurring utterances. In English and Swedish, the notion "nuclear stress" seems to be a useful concept in describing what happens when narrow focus is placed on a particular word or phrase. In both languages, focus is related to the placement of a phrasal tone (the "phrase accent" proper — see Bruce, 1977; Pierrehumbert, 1980) which is associated to the word with nuclear stress. In a context that puts broad focus on a sentence (with no single word or phrase particularly more salient in the discourse), the nuclear stress will be late in the utterance. Narrow focus on a word early in the utterance can be effected by associating the phrase accent to that word, thus effectively displacing the nuclear stress to a word other than the one that would normally bear it. Material after the early focus will then be differentiated from material before it, either by deaccenting all following words in the intermediate phrase in the case of English (Pierrehumbert, 1980), or by downstepping the accents of all following words, in the case of Swedish (Bruce, 1982). In terms of the discourse structure, we can think of this pattern as a strategy of increasing the relative stress of one word by reducing (or removing) the prominence of following accents.

This strategy differs markedly from that in Japanese and other languages that are prosodically like Japanese. In these languages, focus primarily involves patterns of phrasing and there is nothing like the notions "phrase accent" or "nuclear stress". The fundamental frequency modeling work described above shows that both Japanese and French have a tonally demarcated smaller prosodic phrase within the larger intonational constituent, and narrow focus seems to be realized primarily by deleting the tonal marks at following phrase boundaries. In Japanese, the deletion of later tones effectively groups everything after the focused word into a single accentual phrase together with the focused word (Pierrehumbert and Beckman, 1988). In French, on the other hand, the deletion of post-focus tones does not apply to the rise at the end of the focused word itself unless the word is very short. Unlike in Japanese, therefore, the deletion of tones after narrow focus groups the post-focus material separately from the focused word into a kind of postfocus "tail". This is somewhat oversimplified, of course, in that standard European French today is beginning to acquire an "extra" pitch accent — *l'accent d'insistence* — that is inserted toward the beginning of words with narrow focus (Touati, 1987). Still, in both languages the general strategy is one of reducing the salience of following material relative to the word with narrow focus by erasing "normal" tonally-marked minor phrase boundaries, where English (lacking this level of phrasing) reduces the relative salience of following material by erasing "normal" tonally-marked stresses. Thus, marking the edges of minor prosodic phrases in a language like Japanese is some ways functionally equivalent to marking stress by pitch accent placement iin a language like English. The distinction becomes particularly important when we try to extend the investigation of prosody and focus from the constructed material of lab speech to other, richer rhetorical styles or to spontaneous dialogue.

BEYOND SYNTAX

Indeed, it seems fair to say in general that the extensive gathering of production data on fundamental frequency and the modeling of fundamental frequency patterns in association with phonological description of intonational categories has now led us to a point where we are beginning to glean more useful insights into the functions of prosody in natural dialogue. Much earlier work on prosody concentrated on its relationship to syntactic structure. Phonologists have long attempted to predict the stress pattern of an English utterance from its syntactic organization. In the same vein, phoneticians have long investigated the role of prosody in disambiguating syntactically different but segmentally identical strings such as *Fast man offrade bonden, och löparen hälsade kungen* (Bruce, Granström, Gustafson, and House, 1992). Since the occasions must be extremely rare when comprehension hinges crucially on deciding between such contrastive readings, this research may seem irrelevant to the technology of speech synthesis or spoken language understanding. However, it would be a mistake for those concerned with technological applications to dismiss the results of such research. Even in the absence of two or more likely syntactic parses, prosody organizes speech in a way that is apparently critical for understanding. Pitch range relationships and accentuation patterns help listeners to parse topic structure and to resolve anaphoric reference (e.g., Hirschberg and Pierrehumbert, 1986). Work such as that of Silverman (this volume) demonstrates clearly that modeling such aspects of prosody is paramount in achieving natural and easily intelligible synthetic speech. Moreover, given the otherwise contrast-obliterating effect of prosodic position on segmental realization (e.g., Pierrehumbert and Talkin, 1992), a good understanding of intonation and prosody is also directly relevant for robust recognition of segments. The last decade and a half of research on prosody has thus taken us well beyond an inordinate emphasis on syntactic contrasts.

With this coming of age of our understanding of the role of prosody in recognizing segments in connected speech and in cueing discourse structure, there is also a very encouraging merging of research traditions. On the one hand, there is the detailed sorting out of prosodic categories proper in well-controlled phonetic experiments (the many references cited above), and on the other, there is the more impressionistic descriptions of

larger speech corpora, including records of spontaneous speech (e.g., Altenberg, 1987). Researchers trained in one or the other of the sets of disciplines relevant to these two different traditions are now taking better advantage of each other's research in analyzing large corpora. This trend was well illustrated at the last International Conference on Spoken Language Processing by the many papers that combined the two approaches. As it becomes more feasible to gather and store ever larger corpora, the importance of well-controlled background work in the laboratory becomes more apparent. And, as Ostendorf, Price, and Shattuck-Hugnagel (this volume) point out, a good understanding of the relevant prosodic units (such as that which has been achieved for many languages through the detailed laboratory work of the last decade and a half) is prerequisite to the use of more general stochastic models in analyzing large spoken corpora. We see recognition of this especially in the emergence of cooperative efforts to build prosodically labelled databases, such as the development of the ToBI conventions for transcribing intonation and phrasing in English (Silverman et al. 1992).

BEYOND FUNDAMENTAL FREQUENCY

One thing that has also become clear with the development of these prosodic labelling systems, however, is how much room for basic research there still is in aspects of prosody other than fundamental frequency modeling. Work on articulatory correlates of rhythmic structure, such as De Jong (1991), make clear how poorly understood are the phonetic bases of local and global variations in speech timing. An equally large problem is the dearth of basic psychoacoustic research for relevant psychological correlates other than pitch. We know enough about pitch perception now that we can intelligently compare different phonetic representations (e.g., Hermes and van Gestel, 1991). By contrast to this, our understanding of the perception of spectral dynamics is still very limited and new, so new that we have advanced little beyond the guess that durational correlates of stress in languages such as English and Swedish might be related somehow to the temporal summation of loudness (e.g. Beckman, 1986).

However, here again, I see strong grounds for optimism. We are at least seeing renewed attention to aspects of the signal other than fundamental frequency (e.g. Bartkova, Haffner, and Larreur, this volume), and as our understanding of other aspects besides the fundamental frequency patterns improves, so should our prosodic models and synthesis. Also, as our understanding of speech timing and of more subtle spectral cues to voice source patterns improves, we should begin to be able to answer currently puzzling questions concerning the role of these other phonetic dimensions of prosody in differentiating pragmatic interpretations of the same intonation pattern (e.g., Hirschberg and Ward, 1992). Let us hope together for another fifteen years of productive research on these aspects of prosody.

REFERENCES

B. Altenberg (1987), *Prosodic Patterns in Spoken English* (Lund University Press).

M. J. Anderson, J. B. Pierrehumbert, and M. Y. Liberman (1984), "Synthesis by rule of English intonation patterns", *Proc. IEEE Internat. Conf. Acoustics, Speech and Signal Processing*, pp. 2.8.2-2.8.4.

C. Avesani (1990), "A contribution to the synthesis of Italian intonation", *Proc. Internat. Conf. Spoken Language Processing*, Vol. 1, pp. 833-836.

M. E. Beckman (1986) *Stress and Non-Stress Accent* (Foris, Dordrecht).

E. Gårding (1987), "Speech act and tonal pattern in Standard Chinese: constancy and variation", *Phonetica*, Vol. 44, pp. 13-29.

G. Bruce (1977), *Swedish Word Accents in Sentence Perspective* (Gleerup, Lund).

G. Bruce (1982), "Developing the Swedish intonation model", *Working Papers, Department of Linguistics, University of Lund*, No. 22, pp. 51-116.

G. Bruce, B. Granström, K. Gustafson, and D. House (1992), "Aspects of prosodic phrasing in Swedish", *Proc. Internat. Conf. Spoken Language Processing*, Vol. 1, pp. 109-112.

J. Fletcher and E. Vatikiotis-Bateson (1991), "Articulation of prosodic contrasts in French", *Proc. 12th Internat. Cong. Phon. Sc.*, Vol. 4, pp. 18-21.
J. Fokes and Z. S. Bond (1993), "The elusive/illusive syllable", *Phonetica*, Vol. 50, pp. 102-123.
H. Fujisaki, H. Hirose, P. Halle, and H. Lei (1990), "Analysis and modeling of tonal features in polysyllabic words and sentences of Standard Chinese", *Proc.Internat. Conf. Spoken Language Processing*, Vol. 1, pp. 841-844.
H. Fujisaki and H. Sudo (1971), "A generative model for the prosody of connected speech in Japanese", *Ann. Rep.Engineering Research Institute, University of Tokyo*, Vol. 30, pp. 75-80.
C. Gussenhoven and T. Rietveld (1992), "A target-interpolation model for the intonation of Dutch", *Proc. Internat. Conf. Spoken Language Processing*, Vol. 2, pp. 1235-1238.
J. 't Hart and R. Collier (1973), "Intonation by rule: a perceptual quest", *J. Phonetics*, Vol. 1, pp. 309-327.
D. J. Hermes and J. C. van Gestel (1991), "The frequency scale of speech intonation", *J. Acoustical Soc. America*, Vol. 90, pp. 97-102.
J. Hirschberg and J. Pierrehumbert (1986), "The intonational structuring of discourse", *Proc. 24th Assoc. Computational Linguistics*, pp. 136-144.
J. Hirschberg and G. Ward (1992), "The influence of pitch range, duration, amplitude, and spectral features on the interpretation of the rise-fall-rise contour in English", *J. Phonetics*, Vol. 20, 241-251.
K. De Jong (1991), *The Oral Articulation of English Stress Accent*, doctoral dissertation, Ohio State University.
S. Jun (1989), "The accentual pattern and prosody of Chonnam dialect of Korean", in *Harvard Studies in Korean Linguistics*, ed. by S. Kuno et al. (Harvard University Press, Cambridge), pp. 89-100.
K. Kohler (1991), "Prosody in speech synthesis: the interplay between basic research and TTS application", *J. Phonetics*, Vol. 19, pp. 121-138.
D. R. Ladd (1987), "A phonological model of intonation for use in speech synthesis by rule", *Proc. EuroSpeech*, pp. 21-24.
B. Lindblom (1990), "Explaining phonetic variation: a sketch of the H&H thoery", in *Speech Production and Speech Modeling*, ed. by H. J. Hardcastle and A. Marchal (Kluwer, Dordrecht), pp. 403-440.
J. Pierrehumbert (1980), *The Phonetics and Phonology of English Intonation*, doctoral dissertation, MIT.
J. Pierrehumbert and M. E. Beckman (1988), *Japanese Tone Structure* (MIT Press, Cambridge).
J. Pierrehumbert and D. Talkin (1992), "Lenition of /h/ and glottal stop", in *Papers in Laboratory Phonology II: Gesture, Segment, Prosody*, ed. by G. J. Docherty and D. R. Ladd (Cambridge University Press, Cambridge), pp. 90-117.
Y. Sagisaka (1990) "On the prediction of global F0 shape for Japanese text-to-speech", *Proc. Internat. Conf. Acoustics Speech and Signal Processing*, pp. 235-328.
C. Shih (1988), "Tone and intonation in Mandarin", *Working Papers, Cornell Phonetics Laboratory*, No. 3, pp. 83-109.
K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg (1992), "TOBI: a standard for labeling English prosody", *Proc. Internal. Conf. Spoken Language Processing*, Vol. 2, pp. 867-870.
N. Thorsen (1980), "Neutral stress, emphatic stress, and sentence intonation", *Ann. Rep., Institute of Phonetics, University of Copenhagen*, No. 14, pp. 121-205.
P. Touati (1987), *Structures prosodiques du suédois et du français* (Lund U. Press).
N. Willems, R. Collier, and J. 't Hart (1988), "A synthesis scheme for British English intonation", *J. Acoustical Soc. America*, Vol. 84, pp. 1250-1261.