"LINEAR" AND "OVERLAY" DESCRIPTIONS: AN AUTOSEGMENTAL-METRICAL MIDDLE WAY

D. Robert Ladd
Department of Linguistics, Edinburgh University

ABSTRACT

(1) All current descriptions of intonation involve strings of local events - pitch accents and phrase-final pitch movements. (2) Pitch range effects reflecting hierarchical structure affect individual accents, not whole domains. These two facts argue for neither an overlay model nor a radically linear one, but one with a "metrical" structure that specifies the relative height of prosodic constituents.

INTRODUCTION

Overlay vs. Linear?...

Let me begin by focusing in on what I see as the central disagreement at issue in this symposium. It is not, as Beckman's paper [2] reminds us, whether "superpositional" or "overlay" structure exists in F0. Anyone dealing with intonation must inevitably deal with its superpositional aspects. Anatomy (or culture, or both) makes some speakers' voices high and others low, some lively and others flat. Individual speakers may raise their voice or lower it, and may talk animatedly or monotonously. In many languages speakers may aim at a variety of distinctively different pitch levels or patterns, and in all languages (as far as anyone can determine; cf. [20]) there are segmental influences on the fine detail of F0. All of these effects can be fairly readily distinguished, and they interact in a way that is appropriately described as superposition.

I think we all agree on that much; we disagree on two more specific points. First, there is disagreement about which phenomena belong to which superposed layer. For example, are paragraph cues like raising and lowering the voice, or are they like choosing between distinctively different pitch patterns? This issue is unfortunately largely beyond the scope of this paper, though I will return to it tangentially at the end. Second, even if we agree which phenomena belong to the narrowly linguistic system we call intonation, we disagree about whether a superpositional model is appropriate for that system alone. We can all at least agree that intonation includes cues to phrasing and prominence and sentence-type, but some of us (e.g. [3,10,14,17] want to describe these in terms of strings of phonological "events", while others (e.g. [7,8,16,18,19]) want to model them as part of the general superpositional nature of F0. This is the issue I wish to discuss here.

...or Hierarchical plus Linear?

Specifically, I wish to defend a basically linear view of intonational phonology, but one in which hierarchical organisation - tree structure, roughly speaking - plays a significant role. I will focus on two phenomena: (1) mismatches between the phonetic extent of a pitch phenomenon and its functional "domain", and (2) hierarchical phrase-level effects or "paragraph cues". The first of these poses problems for the overlay approach, while the second is particularly difficult to accommodate in the radically linear approach associated with Pierrehumbert and Beckman's work [2,3,4,17]. Unfortunately, neither of these is a straightforward empirical problem, because in some sense F0 is really pretty easy to model: many approaches yield synthetic intonation that sounds fairly acceptable, and by any measure many models are capable of approximating natural data. However, the theoretical arguments can now be based on a much larger body of solid empirical data than was true twenty or even ten years ago.

Again, let me begin with a point of agreement. I believe that at some level of abstraction we are all describing intonation in terms of strings of events, and that therefore in some sense we are all operating with a linear phonology. For example, Grønnum [8,18,19], Möbius [16], and Fujisaki [7] all assume a string of accents, associated with specific syllables in the string of words. In the phonology, these could be represented autosegmentally, and an overlay model of intonational phonetics can be regarded as the "phonetic realisation" of that phonological string.

In saying this, I am not trying to argue that "those overlay people are basically doing what we linear types do". If anything, I acknowledge the possibility that "we linear types are basically doing what those overlay guys do". That is, I am quite prepared to accept that a Fujisaki-style "accent command" is a possible phonetic model of an autosegmental L+H* accent. In fact, the Anderson et al. implementation of the L+H* accent [1] looks remarkably like a Fujisaki-style accent command. My point is simply that all of us are looking at accents in very similar ways, and that one of the fundamental empirical tasks in which we are all engaged is to model *the phonetic details of accents in succession*. Where we disagree is over how to do it.

THE RELEVANCE OF DOMAINS

The basic assumption of the overlay approach is that every prosodic domain has characteristic pitch features, and that these pitch features *extend over the whole domain that they characterise*. Thus accent is a property of words, and dictates the shape of the pitch contour of words. "Declination" is a property of phrases and utterances, and perhaps also of paragraphs as well, and dictates the overall trend of pitch throughout the domain. And lexical tone, in languages that have it, is supposed to be a property of individual syllables.

This assumption is not unreasonable, but it is also not a single assumption. It is possible to imagine pitch features that characterise prosodic domains but do not extend over the whole domain they characterise. For example, we might imagine a sharp local rise marking the end of every intonational phrase, or the beginning of every paragraph, or whatever. That is, even if we admit the existence of a hierarchy of phonological domains - syllables, words, phrases, utterances - it does not follow that each one has its own slope or shape. Conversely, we can imagine pitch features that extend phonetically over something that is not a "domain" at all, but only an essentially accidental stretch of speech from one local event to another. Both these kinds of phenomena exist, and both are problematic for the overlay view.

The Hat Pattern

Taking the second case first, consider the "hat pattern" in the IPO description of Dutch intonation [9]. This consists of a rising pitch movement on one accented word, followed by a high level stretch, followed by a falling pitch movement on the next accented word. (This is, or is similar to, the intonation pattern shown in Möbius's Fig. 4.) In autosegmental or linear terms, the hat pattern is a high (H*) accent followed by a H*+L or H+L* accent, and presents no problems for the theoretical approach.

In an overlay model using Fujisaki's or Möbius's "accent commands", the hat pattern can easily be modelled - *phonetically* - by placing the beginning of the accent command at the first accented word and the end of the accent command at the second accented word. But (as Möbius would be the first to agree) this description makes no sense functionally or phonologically. For Möbius, each accent is the heart of a domain called an "accent group", and each accent group is supposed to have its own accent command. The hat pattern uncontroversially contains two accented words, but has the phonetic shape of a single accent group. This is a paradox for the overlay model.

I emphasise that the issue is not simply the ability to model F0 contours as phonetic objects. That, as I said above, is relatively easy to do (compared to, say, modelling segment duration). The real issue is to relate the parameters of one's phonetic model of F0 to distinct aspects of intonational function - i.e. linguistic categories such as phrase, accent, degree of emphasis, focus, and so on. If "accent group", as a linguistic construct, is a central part of one's phonetic model of accent, then it is a problem if we have to ignore the accent group in order to make the phonetic model "work". (I should add that Möbius is very careful to constrain his use of phonetic paramaters in a way that makes sense linguistically. However, many proponents of overlay models - including Fujisaki - are rather less so.)

Edge Tones

There is also clear empirical evidence of pitch phenomena that relate functionally to a large domain but have localised phonetic manifestations. The best examples are *edge tones* of various kinds. I use "edge tone" as a cover term for what are variously called boundary tones, phrase tones, phrase accents, non-prominence-lending pitch movements, and so on. These have gone from being a source of theoretical bemusement 20 or 30 years ago to being a generally accepted part of the theoretical landscape today.

For example, in their early work on what became the IPO analysis of Dutch intonation, Cohen and 't Hart [6] explicitly pointed out the difference between the fall at the end of a statement and the rise at the end of a question as follows: "This so-called question rise need not occur in dominant words or even on prominent syllables, as opposed to final falls. In other words, this rise should not be taken to replace a final fall, but must be seen as an added feature" (p. 189). In the terms used here, the final fall is a pitch accent while the question rise is an edge tone. In current IPO terms, the final fall is accent-lending and must occur at the stressed syllable of a word that is prominent in the utterance; the question rise occurs at the very end of the utterance irrespective of the location of stress, is non-accent-lending, and must therefore be preceded somewhere in the utterance by an accent-lending pitch movement. Whichever way we express the difference, it is not a difference that would surprise anyone today - but in 1967 it required special comment.

Note that Möbius's model in effect includes edge tones, in the form of phrase commands at the *end* of a phrase. For the most part, phrase commands occur at the beginning of phrases, and serve to model the course of declination. In addition, however, Möbius uses them at phrase ends: a normal phrase command to model a phrase-final rise in pitch (a high boundary tone or IPO type 2 rise), and a *negative* phrase command to model a phrase-final drop in pitch. As I noted above, Möbius insists on some sort of linguistic or functional constraints on the location of the two types of command: accent commands are for phonetic effects related to prominent words, and phrase commands for phonetic effects related to phrases. His use of a phrase command to model phrase-final pitch movements is therefore commendable in principle; the only problem is that, as Liberman and Pierrehumbert have shown [14], the phrase command is really not a very accurate phonetic model of what happens at the ends of phrases. In effect, for phonetic accuracy, Möbius might better use half an accent command to model these boundary phenomena, but this route is closed to him on theoretical grounds.

Significance of Edge Tones

One of the first works to demonstrate the need for edge tones, Gösta Bruce's PhD thesis [5], also makes clear their significance for the overlay-vs.-linear debate. Bruce's specific concern was to develop an account of how lexical accent distinctions in Stockholm Swedish are manifested phonetically in different sentence contexts, but his solution to this problem lays the foundation for a more general theory of how word-level and sentence-level features interact.

In Swedish, the main stressed syllable of each word, in addition to being stressed, bears one of two accents, often called simply Accent 1 and Accent 2. The phonetic difference between the two accents is very striking in some environments and exceedingly subtle in others, but it typically involves a difference in the pitch contours of words and is therefore often described as a difference of "pitch accent". In Stockholm Swedish, the phonetic difference between the two accent types in citation form is superficially a difference between single peaked (Accent 1) and double peaked (Accent 2) pitch contours, as can be seen in Fig. 1.

Bruce's work made clear that the citation form contours involve an interaction between word level accent features and phrase- or sentence-level intonation features. In some sense this was never in doubt, but Bruce's breakthrough was to state explicitly what the accentual and intonational features are. Specifically, he established that the genuinely distinctive feature for the two accent types is the *alignment of an*

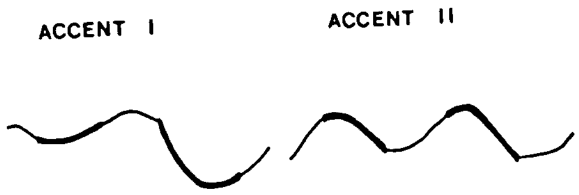


Figure 1. F0 contours of citation forms of "Accent 1" and "Accent 2" in Stockholm Swedish. Adapted from Bruce [5]. These contours are for two-syllable words stressed on the first syllable. The thinner and thicker line sections show the duration of the consonants and vowels respectively.

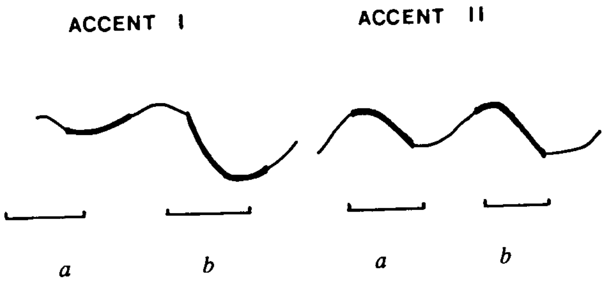


Figure 2. Bruce's analysis of the contours from Fig. 1, showing (a) the different alignment of the accentual fall, and (b) the high-low sequence that signals the end of the phrase and of the utterance.

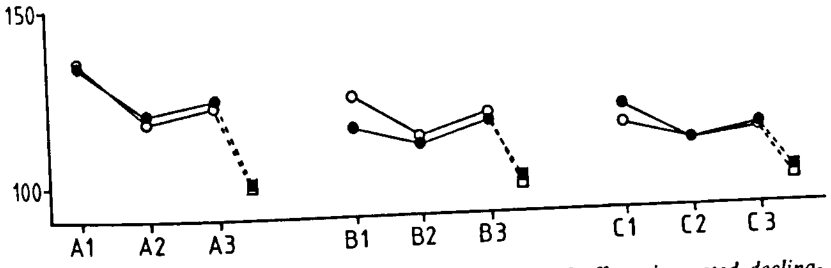


Figure 3. Sample data from Ladd's study of hierarchical effects in nested declination [10]. The circles show the mean F0, for one speaker, of the three accentual peaks in three consecutive clauses A, B, C; the squares show the clause-final low F0 endpoint. Filled circles and squares show data from the A and B but C *structure; open ones show the* A but B and C *structure. For more detail see text.*

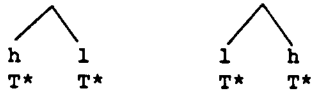
underlying pitch peak with the stressed vowel. In Accent 1 the peak precedes the onset of the stressed vowel by a considerable extent, so that if there are no preceding unstressed syllables the stressed vowel simply begins mid or low, while in Accent 2 the peak and the subsequent drop in pitch more or less coincide in time with the stressed vowel and are therefore always present in the phonetic F0 contour.

The single peaked and double peaked word contours in citation forms result from the interaction of these invariant word accent features with features of sentence intonation. Crucially, these "features of sentence intonation" are not overall trends, but edge tones - a late peak or "phrase accent", and a final fall to the bottom of the range. When these edge tones occur after Accent 2, in which the accent has already produced a clear peak and fall on the accented vowel, the result is a "second" peak. But when they follow Accent 1, in which the accentual high may not be realised as such, the result is a phonetic rise across the accented vowel to the single peak in the utterance. This analysis is shown in Fig. 2.

In addition to providing an elegant and convincing solution for a long-standing problem of Scandinavian phonology, Bruce's analysis clearly shows the nature of the interaction between pitch features whose function relates to domains of different sizes: the word accents and the sentence level intonation features interact not by overlaying small-domain features on large-domain ones, but as a *sequence of phonetic events.* The sentence-level features affecting the word accent contours in citation form are not global shapes or trends but localisable phonetic events. This does not, of course, establish that phonological structure based on superposition is impossible, but it strengthens the argument for a rigorously linear phonological model, because it shows that such a model provides an accurate account of a case that prima facie might be expected to fall within the scope of the overlay approach. That is, if large-domain phonological specifications are unnecessary in these cases, then parsimonious theorising suggests we ought to try to do without them altogether.

HIERARCHICAL PITCH RANGE

The second kind of phenomenon I want to discuss is the manipulation of pitch range to convey the overall organisation of discourses - what we can informally lump together under the heading *paragraph cues.* I have proposed elsewhere [10,11,12,13] that these are properly thought of in terms of *abstract relations of relative height* between prosodic constituents of various sizes. That is, paragraph cues involve phonological relations in a metrical tree or similar structure, the phonological relations being used to express not relative prominence (as in standard metrical phonology, e.g. [15]) but relative pitch range. Thus the distinction between downstepping and non-downstepping accents can be represented as the difference between the following two "metrical" relations:



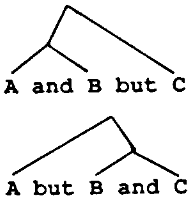
(Here T* is any pitch accent, the h-l relation means the second pitch accent is downstepped, and the l-h relation means it is not.) I have suggested that this approach could be extended to relations between higher level prosodic domains as well, and that the details of the nested relative-height relations could be used to account for detailed differences of relative height. My previous proposals are admittedly pretty sketchy about exactly how hierarchical relations translate into quantitative parameters, a failing I am not going to remedy here.

This metrical/hierarchical approach, it seems to me, falls neatly between the radical linear approach and the overlay approach. It agrees with the overlay approach in potentially being able to accommodate the indefinite nesting of prosodic domains of increasing size (phrase, sentence, paragraph, etc.). But it agrees with the linear approach in conceiving of the whole problem as being one of specifying *local* phonetic properties of individual accents - in this case, specifying the pitch range of accents based on their position in a hierarchical prosodic structure and the details of the pitch range relations specified in that structure - rather than specifying global phonetic properties of whole domains.

I believe that this metrical approach to relative pitch range is superior to either the strictly linear or overlay approaches. Let me briefly try to defend that claim.

The Overlay View

First let me compare the metrical view with the overlay approach. In [10] I investigated the details of "nested downtrends" - declination-within-declination effects of the sort that form an important argument for the overlay view (e.g. [18,19]). Specifically, I compared three-clause sentences of the form *A and B but C* or *A but B and C*, where A, B, and C are the clauses. There was very clear evidence of nested declination in all cases. However, the results also showed that the duration of the pauses between the clauses, and the height of the initial accent peaks in clauses B and C, is affected by the hierarchical organisation: in both structures the *but*-boundary is "stronger" (in the sense of being preceded by a longer pause and followed by a higher accent peak) than the *and*-boundary at the comparable location in the other structure (see Fig. 3). That is, the results seem to reflect a difference of hierarchical structure best represented as follows:



According to the overlay view, we should expect to be able to account for the observed differences in terms of a declination component for the sentence as a whole, one for each clause, and, crucially, one for the intermediate constituent *A and B* or *B and C*. But given the phonetic details of my results - specifically, the fact that the pitch range differences were concentrated on the initial accents of the B and C clauses - I see no ready way of modelling this with simple overlaid components, and no adherent of the overlay approach has come forward since the publication of my study to show how it can be done. I believe that close investigation of similar cases would reveal equally systematic manipulation of the pitch range of *individual* accents, and that the only way to account for these in an overlay model would be to posit implausibly complex and specific long-domain components, with bumps and dips in just the right places.

The Radical Linear View

The case for the radical linear approach to paragraph cues is, if anything, even worse. Proponents of this view have put themselves in the position of having to ignore a great deal of lawful or systematic phonetic variation, by maintaining that it is all gradient and paralinguistic. Specifically, Beckman and Pierrehumbert argue that pitch range effects like those just discussed are the result of individual phrase-by-phrase settings of a gradiently variable parameter that sets the initial pitch range for the phrase as a whole. The variability of this parameter is not phonological, but is said to be paralinguistic and to reflect the newness of the topic.

It is clear from many studies that overall pitch range is a powerful signal of speaker interest or emotional involvement. Extending this paralinguistic function of pitch range into the realm of grammar, Beckman and Pierrehumbert [3] claim that "pitch range is raised when initiating a new topic, and lowered when concluding one". This implies that "for discourse segments consisting of only one topic, a downtrend is accordingly predicted" (p.300). In explaining the data from the *A and B but C* study, they would presumably have to say that the downtrend is moderated at the *but*-boundary because the topic of the clause that follows is new, or at least not quite as old as the topic of the clause that follows an *and*-boundary.

Actual refutation or falsification of such a proposal is difficult. Nearly everyone would agree that paralinguistic factors are involved in intonation somehow, and at our present stage of understanding it is difficult to draw the line between those factors and phenomena that are genuinely linguistic. However, in my view the paralinguistic account of data like the *A and B but C* data is distinctly implausible. There are three reasons for this.

First, the match between structurally

defined boundary strength and pitch range is quite precise, yet the paralinguistic explanation rules out explicit reference to structure. If we claim that the pitch range data depend only on the speaker's estimation of the newness or interest of a particular clause, we must surely also give some explanation for the close coincidence between the structure and the paralinguistic signals of newness.

Second, it is insufficient to come up with a paralinguistic explanation only for pitch range, since the data showed that there were consistent effects on the duration of inter-phrase pauses as well. In my view, a far more plausible explanation is that details of both duration and pitch range are governed by linguistic differences in the hierarchical organisation of prosodic domains.

Finally, as has been demonstrated in Pierrehumbert's own work [17], it is actually quite easy to distinguish quantitatively between uses of pitch range that are unarguably paralinguistic, such as raising the voice for added overall emphasis, and what I am calling relations of relative height. Experiments in which speakers produce specific contour types with different overall pitch ranges show clearly that the constant pitch level proportions of one peak to another are maintained, regardless of the experimental manipulation of overall range. In my view this makes it implausible, *pace* Beckman and Pierrehumbert [4], to lump all these pitch range effects together as "paralinguistic".

Moreover, as I have discussed elsewhere [12], the problems with the Beckman-Pierrehumbert approach go beyond the difficulties of accounting for detailed structural effects in nested declination. An even more serious conceptual difficulty, in my view, is that Beckman and Pierrehumbert model downtrends in two quite different ways. At the lowest level of prosodic organisation - within what they call the "intermediate phrase" - downtrends are the result of phonologically triggered "catathesis" or *downstep*. At all higher levels, downtrends are the result of the paralinguistic signalling of newness, and merely "mimic" phonologically determined downstep. Why one should mimic the other is never made clear; it would seem more appropriate, given two similar phenomena, to try to ascribe them to similar causes.

This is not an isolated problem. There are other apparently distinctive pitch range relations that can hold between prosodic domains of quite different sizes. One is the "answer-background" relation posited by Liberman and Pierrehumbert [14]. This is seen in the relative pitch range on the accent peaks in the sentence *Anna came with Manny* (answering a question such as "What about Manny? Who came with him?"). This pitch range relation can hold between long and complex phrases, as in a possible rendition of the sentence *I'd actually like to stop talking and go out in search of a beer, if only I could get my point across to my fellow panellists.* But for Beckman and Pierrehumbert this latter utterance would involve at least 3 or 4 separate paralinguistic choices of pitch range, one for each intermediate phrase in succession. This view leaves us with no way to express the fact that the pitch range relation between the whole first half of the sentence and the whole second half is intuitively the same at that between *Anna* and *Manny*.

Another distinctive pitch range relation - perhaps it is the same as the preceding one, although it feels different because the lower phrase ends with a low final boundary instead of a high - is used in alternative questions. The alternatives can be as small as a single accent, or as long as an entire complex utterance. An example of two accents related this way is: *Do you want coffee or tea?* An example of two utterances related this way is: *Do you think we ought to leave this debate where it stands right now? Or would it be better to carry on until one side acknowledges that the other is right?* Cases like these are easy to accommodate in a theory in which pitch range relations can (indeed, must) hold between constituents at all levels of the prosodic hierarchy for which pitch features are specified. But in the radical linear view, most of the phenomena just described have to be treated as unsystematic and paralinguistic, and similarities ascribed to "mimicry". This raises the issue of which phenomena belong centrally to intonation; unfortunately, as I said at the beginning of the paper, there is no space to discuss this here.

CONCLUSION

There are no real conclusions yet, and there is space only for an observation. In my view, the most important point to keep in mind as we work toward a resolution of the linear/overlay debate is that the issue is not merely one of phonetic modelling, but of phonetic modelling constrained by assumptions about linguistic structure and function. This means that there are two approaches to evaluating competing models: one is to assess their accuracy as phonetic models, and the other is to examine their linguistic assumptions. The latter approach is the one I have taken here.

REFERENCES

[1] Anderson, M.; Pierrehumbert, J. B.; Liberman, M. (1984). "Synthesis by Rule of English Intonation Patterns." In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing,* 2.8.2-2.8.4. New York: IEEE.

[2] Beckman, M. E. (1995). "Local shapes and global trends". This volume.

[3] Beckman, M. E.; Pierrehumbert, J. B. (1986) "Intonational Structure in English and Japanese." *Phonology Yearbook* 3, 255-310.

[4] Beckman, M. E.; Pierrehumbert, J. B. (1992) "Comments on chapters 14 and 15." In G. J. Docherty and D. R. Ladd (eds.), *Gesture, Segment, Prosody: Papers in Laboratory Phonology II.* Cambridge: Cambridge University Press, pp. 387-397.

[5] Bruce, G. (1977). *Swedish word accents in sentence perspective.* Gleerup, Lund.

[6] Cohen, A.; 't Hart, J. (1967). "On the anatomy of intonation". *Lingua* 19, 177-192.

[7] Fujisaki, H.; Sudo, H. (1971). "A generative model for the prosody of connected speech in Japanese." *Annual Rept. Engineering Resch. Inst., Univ. Tokyo* 30, 75-80.

[8] Grønnum, Nina (1995). "Superposition and subordination in intonation - a non-linear approach". This volume.

[9] 't Hart, J.; Collier, R.; Cohen, A. (1990). *A perceptual study of intonation: An experimental-phonetic approach.* Cambridge: Cambridge University Press.

[10] Ladd, D. R. (1988). "Declination 'reset' and the hierarchical organization of utterances." *JASA* 84, 530-544.

[11] Ladd, D. R. (1990). "Metrical representation of pitch register." In J. Kingston and M. Beckman (eds.), *Papers in Laboratory Phonology I.* Cambridge: Cambridge University Press, pp. 35-57.

[12] Ladd, D. R. (1993). "In defense of a metrical theory of intonational downstep." In H.v.d.Hulst and K.Snider (eds.), *The Representation of Tonal Register,* Dordrecht: Foris Publications, pp. 109-132.

[13] Ladd, D. R. (1993). "Constraints on the gradient variability of pitch range (or) Pitch Level 4 Lives!". In P. Keating (ed.), *Papers in Laboratory Phonology III.* Cambridge: Cambridge University Press, pp. 43-63.

[14] Liberman, M.; Pierrehumbert, J. B. (1984). "Intonational invariance under changes in pitch range and length." In M. Aronoff and R. Oerhle (eds.) *Language sound structure.* MIT Press, Cambridge, pp. 157-233.

[15] Liberman, M.; Prince, A. (1977). "On stress and linguistic rhythm." *Linguistic Inquiry* 8, 249-336.

[16] Möbius, B. (1995). "Components of a quantitative model of German intonation". This volume.

[17] Pierrehumbert, J. B. (1980). *The Phonology and Phonetics of English Intonation.* MIT PhD Thesis, published 1988 by Indiana University Linguistics Club.

[18] Thorsen (Grønnum), N. (1980). "A study of the perception of sentence intonation - evidence from Danish." *JASA* 67, 1014-1030.

[19] Thorsen (Grønnum), N. (1985). "Intonation and text in Standard Danish." *JASA* 77, 1205-1216.

[20] Whalen, D. H.; Levitt, A. G. (1995). "The universality of intrinsic F0 of vowels". To appear in *Journal of Phonetics.*