TOWARDS AN INTEGRATED FORMAL MODEL OF FUNDAMENTAL FREQUENCY IN OVERALL DOWNTRENDS[1]

Firmin Ahoua and David Reid[1]

Abstract

Although there are major differences in the various conceptual models of F_o scaling, we suggest that the corresponding mathematical formulations may be compatible and that the theoretical differences need not hinder the empirical aspects and practical uses of the theories as demonstrated in speech synthesis. The method follows standard practice in Mathematical Logics: combining and "rounding off" the formalisms of the different models, then allowing for a consistent interpretation of the new unified theory. The approach is applied to two current models of decay in intonation curves. The models and then the conflicts between them are described. These latter were used to construct the integrated model. Our short term objective is to validate the application of our approach by testing and implementing empirical instrumental data obtained independently.

1. Introduction.

While a review of the literature concerning F_o scaling (see, for example, Ladd 1984) shows a consensus that in most, if not all, languages speakers tend to lower their pitch (or "fundamental frequency F_o") during parts ("utterances") of declarative sentences, there is no consensus as to which model can best account for this downtrend. When looking for predictability, there is still no agreement as to what part of the F_o contour declines, with respect to what (pitch accent as opposed to time), what constitutes an utterance, what causes the decline, what kind of mathematical

[1]Firmin Ahoua is Professor of Linguistics at the University of Cocody, Ivory Coast.
David Reid is a former Researcher at the Universities of Chicago and Mannheim, U.S.A

formalism to use, what criteria should be used to see if the theory fits the data, or what relation there is between the formalism and the phenomena. It appears worthwhile to look for points of unification or normalisation among the various models and parameters by borrowing some standard concepts of Mathematical Logic.

In general, unless it can be proven that two theories contradict, at least at the formal level as opposed to the conceptual level, one may look for unification or an extension. The modest goal of this paper is to carry out this first step: we indicate one possible approach for showing that two theories are relatively consistent with one another, illustrating this with the two leading models of F_0 scaling, presently at loggerheads. In our exposition we use the word "model" as it is loosely deployed in the literature, namely to refer to a body of explanation for some phenomena. However, we insist on the following important distinctions, which continue throughout the paper. Two separate aspects of each model are considered: its mathematical formalism (formulae, etc.: the model's theory), on the one hand, and the values assigned to the formalism (the theory's interpretation) on the other. A theory can have many different valid interpretations which may even contradict one another. However, the central point is that if a theory has at least one interpretation that makes sense, then the theory is consistent. Therefore, to show that two theories are consistent with one another, we combine them into one theory (carefully redefining the domains of relations, renaming variables appropriately, and whenever possible, adjusting the arguments of relations to facilitate comparison) and find a reasonable interpretation. The resulting model, or modified Fujisaki's model, as we shall for exposition purposes refer to it, exists for the sole purpose of adjusting possible consistent parts of models relative to one another.

In an application of these principles, it behoves us to look at the mathematical bases of established models. Thus the fine points of phonological models do not receive more attention here than is necessary (see Ladd (1984) for discussion); the reader should be warned not to look for extensive discussion of the conceptual differences between the models, nor any discussion of phonological categories and hierarchies.

The two models which we have chosen as the object of our investigation are that of Liberman and Pierrehumbert (1984), and that of 't Hart and Cohen (1973) (further elaborated in 't Hart et al (1982)). (We abbreviate the former model as "LP" and the latter as "'tH", using neutral pronouns to refer to them). Contrary to LP, who use a linear scale, 'tH apply a logarithmic scale. The latter refer to speech movements (instead of speech targets), so that rises and falls occur. The major component of 'tH's theory is declination. This is a global outlook. Declination refers specifically to the trend of the top and bottom lines that define the limits of the local pitch movements. The main construct of this model is the 'hat patterns', the 'pointed hat", and the "flat hat". LP's major component is the downstep factor, a coefficient that lowers the discrete targets.

The organisation of the paper is as follows: in Section 2 we describe these two models as well as sketch some of the elements of model Fujisaki (1981) (henceforth "Fujisaki") relevant for expository reasons. Here we shall assume that the reader is familiar with the mathematical formalism used (primarily elementary operations with logarithms. see also Ahoua 1990). In Section 3 we point out the major points of conflicts between the two models, 'tH and LP. The point of departure in Section 4 is a model in its first stages of development in the literature and known as "resetting". There is not sufficient data or theory of this concept available, but we propose some extensions of it in its present form. In Section 5 we construct our conciliatory model on the basis of the theory of the two warring models considered. Theory is first developed, followed by the interpretation. Section 6 them shows how our model takes into account the difficulties outlined in Section 3.

2. Overall Falls in Phonetic Models.

When the frequencies (F_o) of the voice in an utterance are graphed as a function of time (ignoring long pauses, as in a voice-activated recording), the result often resembles a tilted sinusoidal curve with decreasing amplitude (with the occasional deviation), the tilt being in the sense of a negative slope in most declarative sentences; i.e., when one connects the local maxima ("peaks" or "highs") in these cases they yield a decay of Fo with time or as a relation between peaks, as does connecting the local minima ("valleys" or "lows"). (For illustrative graphs see almost any of the articles cited.) In addition, comparing several such graphs of the same speaker shows that there is a frequency below which the speaker never descends in ordinary speech. That much is clear. It is at this point that the divergences mentioned earlier begin.

Ideally one would start with principles already well established. For example, as with any physical action, it is a tautology to say that physiological factors play some kind of role. Indeed, several connections have been established, although how much is cause, and how much is effect, is another matter. In any case, one cannot simply assert that involuntary physiological factors are the major cause of this decay (see Collier, 1975 and Collier & Gelfer, 1984). In the other direction, in attempting to work backwards from the data to the causes, one notices that the decay appears to be exponential, so a natural procedure is to graph the logarithms of the frequencies as a function of time, and regard the result as is done in 't H, by expressing the pitch in semitones. That is, if p is a frequency, and m is some fixed base frequency, both in Hertz, then 12*log2(p/m) is the equivalent of p in semitones (with respect to m). 't H then finds a linear graph on this scale to match his data. 't H's formula is as follows:

(1) F(t) = D*t + F(0) ("*" signifies multiplication)

whereby t is time (in seconds, minus pauses longer than a quarter of a second); F(t) is the pitch in semitones with respect to a base frequency m, and F(0) the beginning pitch of the utterance. (We have substituted the "F" for 't H's "P" so as not to confound it with LP's "P".) D is the slope of the lines, and is language-specific. For Dutch, for example, 't Hart et al. (1982:143) find :

(2) D = -11/(1.5sec. + ts) for ts < 5 sec.

= -8.5/ts for ts > 5 sec.

whereby ts = the length of the utterance in seconds.

The D applies to three lines: a high line connecting the peaks, a low line connecting the valleys, and a middle line where the pitch sometimes rests when changing between these two extremes. (i.e., mathematically speaking, the lines join respectively the relative maxima, the relative minima, and, roughly, the inflection points.) The separation of the lines is language-specific. Since D has a negative slope, 'tH terms this "declination". See appendice 1 for illustration.

One can also see the overall decay of a sentence as being composed of sections, each representing its own autonomous decay. This is the viewpoint of LP, wherein a theory of "downstep" is explained. LP's model has been the most influential theory of intonation over the last decades and is sometimes labeled as the Tone Sequence Model as its primitives are tones that are phonetically interpreted as Fundamental frequency values. LP's model can be understood as a sequence of discrete tonal events that constitute the overall contour. Downstep is the basic component of that model and is a term that is independently motivated in African languages. Its phonetic manifestation is triggered by a certain sequence of tones (HL), and specifically in English by a H-accent followed by a Low tone (H*+L ...!H*). The effect is that a following High tone of an accent syllable will be lowered or downstepped. Notice, however, that the Low tone may be floating and trigger the downstep of the following High tone (H*..*!H) .

To outline this process we must first distinguish between different types of peak accents. The latter correspond to peaks which are pushed higher by emphasis; i.e., higher than those levels achieved in the same utterance spoken with normal accent; "strongest" here refers to the local property of comparative strength. If a stronger accent is considered together with weaker (pitch) accents immediately preceding it, then this is considered an AB pattern; if it is considered together with weaker accents which follow it, then we have a BA pattern. In the latter case, according to LP, the stronger accent governs the fall of the succeeding peak accents; that

is, the influence is hereditary: the first peak accent influences the second, the second the third, etc. Each such pair is termed a "step accent", since taken together they form a cascading series. Finally, in LP the set of pitch accents is the union of the set of peak accents with the set of step accents. (In 'tH on the other hand, pitch accents consist of only the peak accents.)

LP formulates mathematical relations for this cascading series. We reformulate these a little for purposes of our exposition, but the essence remains the same. For example, LP's rules cover both AB and BA cases, but here we shall break the two cases up.

The Case of BA Patterns

We consider a series of step accents, where the frequencies are expressed in Hertz, or s-1 . The F_0 value of the initial peak accent is labeled P(0) (or, for simplicity, P_0). P(i + 1) is the F_0 value of the peak accent following the peak corresponding to P(i). Then, according to LP (see Liberman & Pierrehumbert (1984: 193); See also appendix 2):

(3) P (i + 1) - r = s*(P(i) - r)

s is a speaker-specific "downstep constant", and r is an utterance-specific "reference value", calculated by:

(4) r = c*(P_0 - b)a + b + d

whereby "b" is the minimal frequency possible for that speaker, and the other constants are other speaker-specific constants. [We have renamed LP's constants "f" and "e", as "c" and "a" respectively here, in order to avoid confusion with symbols used elsewhere.]. Translating the recursive formula (3) into polynomial form,

(5) P(n) = Sn * (P_0 - r) + r

The "reference level" P_0 = r is explained as a non-zero asymtote to which the decaying peaks tend on a graph of frequency vs. "peak number", i.e., the first, second, third, etc. peak.

We note that we have used "Model 1" from LP. A "Model 2" is also briefly mentioned (see Liberman & Pierrehumbert (1984: 207)); its primary difference to Model 1 is to replace Eq.(5) by:

(6) $P(n) = hn^* P_0 + n^*c$ whereby $h = (r-b)(1-s)$, and $c = (1-h)^*b$

We shall continue, however, to equate Model 1 with LP, since this is the model to which this paper pays primary attention.

The Case of AB Patterns

LP replaces s in Eq.(3) by k (roughly 1/s), which again is speaker-specific, reading p(j + 1) as the value of the highest peak accent, and P(j) as the pitch of the previous peak accent. Yet another part of LP's model is "final lowering". That is, in the last pitch accent of the utterance, the pitch is lower than would be expected by Equations (3) or (5), so that replacing the term sn in equation (3) by the term s*l, or equivalently the term sn in equation (5) by sn*l for the last peak accent provides us with that value. l is speaker-specific.

In LP the formulae are based on curve-fitting, though by using phonological criteria and minimal categories. The best model, however, not only assigns a meaning to relations, but to its arguments (the variables and constants) as well, so that the combination of individual meanings matches the meaning of the combination. One such attempt that met with a certain amount of success is presented in Fujisaki (1981). We outline this approach below in order to indicate its compatibility as well to our Integrated Model.. For ease of exposition and later development, we have submitted his formulae to quite a bit of renaming of variables and some algebraic manipulation. Therefore the reader may note a contrast with the formulation as originally presented in Fujisaki.

In brief, in Fujisaki the Fo contour of a sentence is made up by summing up, over the length of the sentence, the non-overlapping Fo contours of the individual words (his positive accent commands, LP's High pitch accents as opposed to Low pitch accents). Each of these individual contours is the addition of two component curves, the "phrase command" and an

"accent command". Each is activated by different signals. The duration between each phrase command signal is assumed equal to the others. In simpler terms, Fujisaki's model is a kind of overlay or superpositional model that contains a global trend and local F_0 movements. His model is a quantitative model for speech analysis and synthesis. The phrase component is modeled as an impulse response: graphically, it rises rapidly to a peak and then decays exponentially towards an asymptote. The accent component is modelled as a step function, which creates a string of steps up and steps down that represent the local rises and falls of pitch at accented syllables. The step function is smoothed by the addition of a time constant and is then added to the phrase component to create the contour (Ladd 1996:25).

In order to clarify our discussion of Fujisaki, we introduce some definitions and remarks at this point:

a. Let A(t, i) and B (t, j) be functions that vary discretely in time (minus pause) at the i^{th} phrase (or tone) command and the j^{th} word accent, respectively. The values of these functions will then indicate the strength of the respective signals at the beginning of their associated intervals.

b. The accent commands will start later and end earlier than the associated phrase command which forms their base. For a given moment in time t we let:

t^{pb} = the duration since the beginning of the phrase command,

t^{pe} = the duration until the end of the phrase command

t^{ab} = the duration since the beginning of the accent command,

t^{ae} = the duration until the end of the accent command

c. The horizontal axis of Fujisaki's graphs is time, and the vertical axis is in terms of $\ln(F_o)$. By a minimal amount of algebraic manipulation, his formulae can also be expressed by using $\ln(F_o/b)$ [which is merely equal to the semitones times ln2/l2], and b is the minimum possible frequency,

as above. (The reader has perhaps already guessed that this formulation can help with a comparison with ’tH.)

d. Let g (t, q) and h (t, r) be Fujisaki’s exponential expressions in terms of time with parameters q and r respectively, which are speaker-specific constants. More specifically, these are functions, one of whose arguments is the speaker, since the other arguments are unknown and the values are very nearly constant. Fujisaki takes them to be equivalent to speaker-specific parameters, having noted this distinction in his experimental approach as well as his conceptual exposition. The utterance component is the concatenation of the individual components of the utterance considered, each phrase component being expressed by:

(7) $\mathbf{A\,(t,\,i)*[g\,(t_{pb},\,q) - g\,(t_{pe},\,q)]}$

Each accent component is formed by:

(8) $\mathbf{B(t,\,j)*[h(t_{ab},\,r) - h(t_{ae},\,r)]}$

(see Fujisaki (1982: 7).)

To arrive at the Fo contour for the utterance, in terms of ln(Fo/b) one sums up the phrase components and the accent components over time and individual accents, i.e., summing the two over t, i, and j. The (very flexible) exponential functions g and h [g(t, q) = q*t*exp(-q*t) and h(t, r) = (1 - (1 + r*t)*exp(- r*t), resp.] are based partly on curve-fitting, but also partly on certain physiological connections, such as subglottal pressure. This pressure on the lung cavity seems to vary with time grosso modo in the same way that frequency does with time, as do certain other physiological mechanisms. This does not necessarily point to a direct cause¬-and-effect relationship, but rather to the fact that these various mechanisms respond to the same stimuli (such as, perhaps, the larynx). (Besides Fujisaki see also Collier 1975.)

3. Conflicts

Each of models 'tH and LP has its weaknesses and its strengths. These come out most clearly when a strength of one is contrasted with a weakness of the other. Thus, keeping in mind the definitions and formulae of these two models as presented in the last section, we present seven areas of conflict. Doing so will guide us in the development of a model designed to resolve these conflicts.

Conflict 1. Look-ahead.

The Eq.'s (1) and (2) of 'tH depend upon tg, the length of the utterance. This assumes that, in order to choose the beginning pitch and the slope of declination, the speaker must know the duration of his utterance in advance. This idea, especially for longer or spontaneous utterances, is disputed by LP and others, despite the general impression to the contrary. As Ladd (1996: 29-30) expresses it, agreeing with LP (1984:220 ff.): "This degree of lookahead may be psycholinguistically implausible, and for speech synthesis models is certainly computationally expensive."

Conflict 2. Reference levels.

Comparing Eq.'s (1) and (2) with Eq.'s (4) and (5) may not immediately show a conflict unless one notes that the value r prevents the decay from proceeding below the asymptote $F_o = r$, the reference level, whereas no such mechanism is present in 'tH. Referring to the line $F_o = b$ as the "baseline", LP remark:

> "Among authors who have tried to be explicit about how F_o contours are scales, log transforms are popular. `t Hart and Cohen (1973) propose a model of Dutch intonation. Such models make wrong predictions because they lack any counterpart in our reference level, which changes with overall pitch range while leaving the baseline (seen in the final L(ow) tones) invariant." (Liberman and Pierrehumbert 1984: 225).

The paper does not elaborate on which kinds of "wrong predictions" are meant, but one possibility is that these refer to predictions obtainable by extrapolation of the curves in order to determine what would happen if the

utterance continued longer than intended. That is, in this case the speaker might be forced to go below the lowest possible frequency: a contradiction. This weakness of 'tH is avoided by LP via the reference value r, but this then points out a weakness of LP: what is the interpretation of r ?

Conflict 3. Ad hoc variables and formulae.

In LP the ad hoc nature of the means of calculating the reference level is admitted:

> …perhaps only the portion of P_0 above b should be relevant in determining r. Except for b, which is identified with the speaker's invariant final low F_0 value, none of the parameters in this equation have any clear interpretation. The parameter d is the translation of 'somewhat', while e and f are just a way of getting a curved function with a minimum number of additional parameters (Liberman and Pierrehumbert 1984: 205).

We may add that the non-linear part of the formula differs from one utterance to another about as much as the value of d, thus emphasizing the ad hoc aspect of this part of the model.

Conflict 4. Pitch range.

Here we repeat the point indirectly made in the quotation cited in Conflict 1: that 'tH does not take into account the pitch range variations. Eq. (4) has already shown us how the reference value depends on the pitch range; also, for LP the pitch range is tied to emphasis, apparently not taken into account in 'tH. (A note of caution when comparing the pitch ranges in the two models: LP measures pitch range from the highest peak, whereas 'tH measures it from the (lower) beginning frequency.)

Conflict 5. Peaks vs time.

In the above equations, even more striking than the difference in vertical axes as treated in the last two conflicts is the difference in horizontal axes: LP referring to peak number, and 'tH using time (minus long pauses). Admittedly, LP thereby avoids the difficulty of "look-ahead" as

explained above, but the idea of eliminating all dependence on time seems a little strange.

Conflict 6. Non-declining patterns.

Despite the differences in axes, some characteristics can be compared. For example, a falling curve will fall in both frames of reference. But whereas for 't*H* all declarative sentences decline, for LP this is far from the case: only those sections after a strong accent decay; just beforehand there is an upstep, and otherwise the line connecting the peaks may be flat or v-shaped (two roughly equal pitches with a lower pitch in between) or follow other patterns than declination would allow.

Conflict 7. Final lowering.

In section 2 we remarked that the highs and lows come progressively closer to one another. The precise manner in which this occurs is represented in 't*H* by the theory in which the low-line is parallel to (has the same slope D as) the top line when on a logarithmic scale (such as semitones). (A lower line on such a scale will, when converted to a linear scale such as Hertz, descend less rapidly.) The constraint that every utterance ends on this line, i.e. as a low, corresponds in this model to the lowering of pitch which one uses to signal the end of an utterance. LP, on the other hand, pays little attention to patterns of lows other than establishing the minial frequency b, and making a few speculative comments (these latter in Liberman and Pierrehumbert (1984): 218 - 219). For LP, then, this pitch lowering is to be found in an unusual lowering of the last high. For 't*H* the last high follows the same pattern as the others.

4. Resetting

The above conflicts seem at first glance to doom any attempt at a formal or mathematical unification. However, upon closer inspection they turn out to be differences in interpretation, including differences in domains (i.e., sets or classes of possible values for the respective symbols); a unification of theories thus implies a union of their respective domains, and often, as here, a new domain extending this union, although one reasonable extension will serve as well as another to show the compatibility of the

original domains. The extension chosen here, that of greater possibility for variability, including discontinuous variability, is inspired by the requirements posed by the solutions needed for the conflicts enumerated above.

This present section is devoted to those components. The central idea is that of resetting, inspired by techniques already known under that name (see, for example, Ladd (1988) and Ladd (forthcoming)). In our explanation of the term, we shall use "strategy", by which we mean the activation of motor nerves, based on predictions, allowing one to adjust one's physiological mechanism to be ready for a longer or shorter utterance. (Of course, if the semantic strategy changes, so do the syntactic and thus the physiological ones, so the term may be a little loose without causing undue confusion.) The motivation is that the curves described by models of decay are dependent upon the strategy of the speaker. Over longer sentences, speakers obviously change strategies as they speak: the curve changes accordingly, not only in shape but also in position (i. e., a different frequency). A function describing this process may very well include arguments of a syntactical nature. For example, Ladd (forthcoming) finds that the amount of resetting is greater for the conjunction "but" than for the conjunction "and".

Moreover, this research established that if the continuous (i.e., non-reset) pieces of a discontinuous (i.e., reset), F_0 contour are compared, then a difference exists between the corresponding points on each piece (e.g., the beginnings) and the syntactical points of the piece. However, this can be difficult, since the resetting is not always obvious as, for example, in a parenthetical phrase. When the resetting is sufficiently frequent, it may manifest itself as a different decay rate. As an example, Umeda (1982) remarked that declination is different according to whether the text spoken was a read list, a read text, or normal conversation. Investigations by Cooper & Sorenson (1977) were not conclusive as to whether resetting occurred, as is pointed out in the article, even though an interpretation of "local inflections" is preferred there. Obviously, further quantitative research is necessary before adequate mathematical functions describing resetting and declination-of-reset-decays can be established. Nonetheless, we assume that such functions exist. A further complication in searching

for such functions is that each function may be composed of several parts: for example, not only may the pulse within a frame of reference be reset, but the frame of reference, asymptotes, baselines, parameters, may all be subject to resetting, and sometimes these are interconnected; for example, resetting a parameter may change the frame of reference, and vice-versa, although neither direction is automatic. We give some examples of these in the next section, but here we wish to say a few words about frames of reference and related ideas relevant to our later discussion.

By a "frame of reference" is meant a system in which something is expressed. Graphically, the axes, together with the means of graphing, form the frame of reference for a curve. Not every baseline or asymptote forms a new frame of reference. For example, the "guidelines" of the connected highs and lows, respectively, do not form one for the F_0 contour unless the curve is expressed strictly in terms of them. In changing the vertical axis from Hertz to semitones, 'tH has changed the frame of reference; LP could have changed (but didn't) the frame of reference to express the results in "pitch above reference level" (F_0 - r). However, these changes of the vertical axes were for convenience of expression, having no correspondence in the interpretations. In Fujisaki, on the other hand, the interpretation seems to be clear that the accent component is based on the utterance component, leading one to feel that the latter forms a new form of reference for the accent component. In this case, however, there is no correspondence for this interpretation in the theory: he combines the function by a simple addition.

5. Proposal for an Integrated Model

In this section we form our model by outlining first the extension of the theories of LP and 'tH and then giving an interpretation. Our integrated model is formed by first defining the symbols, and then the equations, and finally the interpretation. Although this separation requires a bit of patience, we hope that the reader will bear with us. The model is formed as follows:

(a) We take all the symbols of ´tH and of LP, and rename them appropriately (i.e., if two symbols are to always have the same

interpretation, then they are to be assigned the same symbol, and only then).

(b) Strictly, we should introduce new symbols to distinguish them from those of the other models, while keeping enough similarity in the new symbols to remind us of their connection with the old. For example, we could use m^ rather than m, b^ rather than b, etc. But except for some points below that need to be made explicit, this would be awkward, and so we shall trust the reader to make the necessary distinctions according to the context. As well, we could introduce a function to account for the lack of perfect predictive power of the model due to hidden variables, but this may be left implicit:

(i) the variables b^, r^, m^, s^, m^, t^s, F(ô) and Pô, so named as to be associated with the parameters without the hat,

(ii) the variable symbols "state", "syntax", and "Language", and the function D´ (Language, ts),

(iii) for every variable symbol x, introduce a function symbol fx.

For every discontinuity of this function, define another function equalling fx restricted to a neighbourhood of the discontinuity. The union of these latter functions forms a resetting function equalling fx restricted to neighbourhoods of its discontinuities.

(iv) for each function f(x) with parameter p being one of those mentioned in (i), introduce the function f(x, p^).

(v) for every function f, introduce the set of functions (f1, f2, f3,…) so that f may be expressed as their combination (addition, composition, etc…)

(vi) for every function symbol g, introduce the symbol g^.

(viii) for every relation or function symbol R, introduce the symbol DOMR. (In discussion the subscript will be left out if the context is clear.)

(c) Introduce the following new relationships:

(i) In equations (1) - (5) replace m, b, r, D, F(0) and Po by m^, b^, r^, s^, D', F^(0), and p^o^ respectively, and adjust the function symbols as in (b)(v).

(ii) For each variable x introduce the relation:

(9) **x^ = fx(state, syntax, Language)**

(iii) For each application of step (i) transforming an old formula f into a new formula g and transforming y to y^, both with variable x, solve g(x, y^) = f(x) for y^, stating its validity DOM. Example: take Eq. (5) and its transformed version, and we arrive at:

(10) **s^ = (sn*(Po - r) + r - r^]/(Po - r) 01/n**

(This can of course then be combined with (ii); i.e., setting the right-hand sides equal when left hand sides are equal.)

(iv) for every function g, introduce the relation:

(11) **g^ (u(x, state, syntax)) = u(g(x), state, syntax)**

The interpretation of the integrated model is formulated as follows (The upper case letters correspond to the lower case ones of the syntax.):

(A) Keep those symbols which already have interpretations in LP and 'tH; the interpretations of the others will be explained below as special cases of new variables.

(B)

(I) b^ is a base frequency below which the speaker could not descend at a given moment (i.e., given conditions) if he wished to lower his frequency. Thus when the interpretation of "syntax" (see below) includes being at the end of a declarative sentence, b^ = b.

For the next two variables we assume that, among the characteristics of pitch which the speaker uses to modulate his voice, there are bases to which the speaker compares his voice. Among these there would be a (variable) highest and a (variable) lowest value at any point. These are the interpretations of r^ and m^, respectively. It is possible that m^ = b^, but this

requires further research. Fujisaki assumes that m^ = b^ = b, assumptions that we are not willing to make. We assume rather the relationships as in (c) (ii) above; that they depend on "state" and "syntax" (subglottal pressure, anger, emphasis, etc., as below). Furthermore, by "bases" is meant an interpretation corresponding to composition of functions in the theory. s^ has the same meaning as s, except that it is variable as indicated by Eq.(9). The remaining hatted variables are explained in (B)(IV) below.

(II) "state" is interpreted as the set of relevant measurable physical, physiological, and psychological constraints of the speaker at the time considered. "Syntax" is interpreted as the union of (1) syntactical constraints imposed by the language in the context of the speaker's state and (2) relevant previous values of "syntax". (Thus its fuller description would be recursive.)

"Language" is Dutch, English, Japanese, etc. and would include cultural as well as grammatical relations and variables. "Strategy" could also be defined in terms of these three - not necessarily independent - variables.

(III) The replacement of D by D' is necessary since the slope formula is perhaps language-specific or situation-dependent.

(IV) Resetting corresponds to the resetting functions above.

However, we do not use them to define utterances: in the integrated model, an utterance can be any part of a sentence, spoken or intended. Furthermore, f/s is to be interpreted as the duration of an utterance, not necessarily, but possibly of a sentence (the duration of an utterance or possibly a sentence ?). Thus Po^ and F(0) correspond to Po and F(0) respectively, but are the beginnings of the utterances considered, of which there can be many in a sentence, and thus are variable.

(V) This definition of new functions is obviously a formal necessity stemming from the use (explained in (C) below) of the newly defined variables.

(VI) The decomposition of functions allows one to take into account that

more than one function at a time may be relevant: that is, more than one strategy can occur at once, so that the end effect is a combination of the individual corresponding functions. This corresponds to the fact that a human thinks in parallel terms rather than serially, even though the end result is required to be a serial representation. The combination may be an addition of weighted values, as in Fujisaki (the functions A and B providing the weights), a composition of functions (as in the changing of frames of reference), or other combinations of operations.

(VII) The formal necessity of g^ is made evident in Eq.(11).

(VIII) "DOM" indicates the domains of relations, and thus is used to hold in check any unfounded universality.

(C)

(I) The new interpretations of the equations of LP and 'tH, as formally adjusted, follow automatically from the interpretation of the variables above.

(II) Eq.(9) indicates the dependence of the new variables on more fundamental factors. An example of the use of this equation was already given above in mentioning a constraint on b^.

(III) Eq.'s (1) - (5) we take to be valid in the sense that they yield correct data (within fuzziness) under their respective original domains. Thus we constrain our variables to be equal to the values found for the parameters of these equations. The example of b^ was just given; Eq.(10) is an example with s^. In the latter, s^ will be equal to s under the original DOM despite the "hatting" of the other quantities; however, under an extension of DOM, there is no reason to expect it to take on the same values elsewhere.

DOM can here be given a wider interpretation, but only very cautiously. Upon an extension of DOM, the original formulae will likely be shown to be approximations to other formulae, so it is dangerous to claim any universality for a formula developed from limited data. To take a simple example: if, for an utterance that began at 119 Hz and lasted 4 seconds, one calculated the middle line in the model of 'tH, then one could just as

well have used the linear formula F0 = 119 - 11*t, coming within 3 Hertz of the values given by the more complicated formulae (1) and (2) above. Only by observing other cases would it become apparent that the linear formula would not suffice. Thus by expanding DOM one also corrects formulae. Another advantage of such an expansion would be to clarify the resetting functions, as can be seen by combining (c)(iii).

(IV) Eq. (11) is a formal necessity, given our previous definitions.

6. Resolution of Conflicts.

With regard to the difference in approaches, we share the outlook of Beckman and Pierrehumbert (1986: 302) that a theory of downstep (renamed "catathesis" in this later work) need not contradict a theory of declination. Indeed, we show how the seven conflicts of Section 3 are to be resolved in the integrated model, thus reinforcing the idea that the theories, if not the interpretations, of LP and 'tH are consistent with one another. The numbering of the seven resolutions below corresponds to that of the conflicts above.

Resolution 1. Look-ahead.

That the declination of the speaker's voice will vary according to the length of the utterance does not necessarily imply a "look-ahead" strategy, since an alternative interpretation could be that the speaker selects the slope and beginning pitch, thus determining the length of the sentence. However, this latter is somewhat unnatural, so we shall assume the former. The resolution lies rather with our interpretation of "utterance" as any part of a sentence, spoken or only intended. That is, there is no contradiction in assuming that the speaker combines a "look-ahead" strategy with a "see-as-you-go" one in that he makes tentative predictions upon which he acts (speaks) until he makes another one, whereby a "prediction" refers to a physiological and psychological state, not to a conscious calculation. If the speaker changes his strategy, then this would be equivalent to either (1) choosing from a family of states, i.e., dispositions to a new slope and starting frequency, or (2) changing the (physiological, psychological, and

mathematical) frame of reference, and either also changing the slope and frequency, or keeping one or both in the frame. In any case the contour is reset, and the necessity for a correct prediction of utterance duration is avoided.

Resolution 2. Reference levels.

LP praises its reference level as an asymptote for the contour, and criticises 'tH for not having one. Does 'tH need one ? The straight lines in the graphs of 'tH are each specifically defined for a limited domain. That is, the slope of the line is defined by the length of the utterance, and not beyond. In other words, a strategy is identified with a slope and, by extension, with a starting frequency, a frame of reference, base line values, and so forth, over the interval of time corresponding to the duration of the (possibly interrupted) strategy. Reasons for changes in strategy and hence the associated values are multiple, including physiological, psychological, and syntactic grounds. Such changes are thus associated with an adjustment, or "resetting", of these values. Each "reset" then defines a new domain. To talk of extrapolation in this context is inappropriate; asymtotes are unnecessary (an asymptote being a linear extension of an extrapolation to infinity), given proper attention to DOM, to insure that the decay includes the function of "brake" which is performed in LP by the reference level.

Resolution 3. Ad hoc variables and formulae.

A certain amount of curve-fitting is inevitable, but a basic requirement for any model is for the interpretation of the combined symbols to be the same as their combined interpretation. This principle of strict correspondence between theory and interpretation cannot be fulfilled if the individual symbols do not each have an interpretation; however, the workability of the formulae as a whole indicate that with a bit of care, an interpretation could be given to the individual symbols so as to fit this criterion. This was the motivation behind our definitions and interpretations in the integrated model. The interpretations may not be to everyone's taste, but at least there is a correspondence.

Resolution 4. Pitch range.

We return to LP's assertion that a "reference level" is meritorious because it links pitch range with the type of decay (as is seen by combining Eq.'s (3) and (4) or, equivalently, (4) and (5). This link is present as well in `tH. Purely formally, combining Eq.'s (1) and (2) shows that the pitch range varies with the length of the sentence, or, put another way, the length of the sentence varies with the range, until a maximum pitch range is attained. (Although LP does not handle longer utterances, one would presume that it admits the limitations of pitch range.) Even when the pitch range is the same in terms of semitones, it will not be the same in terms of Hertz if the beginning frequencies aren't. Similarly, for the slope, although the semitone scale is actually pushed slightly down by emphasis (emphasis increases utterance duration slightly), the higher beginning frequency would cause a steeper decline on the Hertz scale, as is the case in LP. For `tH, then, the ability of its graphs to take in differing pitch ranges due to emphasis is clear. (In the integrated model, furthermore, a more subtle distinction is made: it is possible for m^ to be reset, altering yet again the relationship between the Hertz and the semitones, just as resetting b^ or r^ can alter the interpretation of LP's theory. Thus this additional flexibility of the integrated model makes contradiction even more unlikely).

Resolution 5. Peak number vs. time.

We continue our assertion that flexibility regarding one or the other frame of reference resolves many a difficulty. The difference in horizontal axes reflects not only two different attempts to fit data to a curve, but also a difference in outlook regarding phonological phenomena: in considering peak number, LP concentrates on the discrete characteristics, whereas 'tH rather considers the continuous aspect. As both aspects are likely in play during an utterance, our model uses one when considering peak number, and another when considering time. Is there a danger of contradiction here? A direct comparison is impossible due to the discrepancy of axes.

Let us illustrate how the two may be compatible by assuming, for the moment, that both are used to analyse an utterance with the first word

heavily emphasized (the so-called "BA" pattern in LP), and where no resetting occurs. Furthermore we shall, in order to make the demonstration simpler, assume that b^ = b, s = s, etc., as well as leaving out the final lowering for the moment (returning shortly to the latter). Given a fixed pair of beginning frequencies (beginning the utterance and the first accented peak), a fixed speaker, and a fixed peak number, is there a time (within reason) when `tH's value = that of LP ? If the answer is in the affirmative for every n before resetting might occur, then the functions are compatible.

Using the notation of section 2, our above question then becomes whether there is, given fixed values for **s, r, b, t_s, Q(0), P(0)** and n, a reasonable value of **t** so that **P(n)** = **Q(t)**, whereby we let **Q(t)** mean **F(t)** converted into Hertz. For the sake of illustration we take the following random possible values:

(12)

s = 0.6

r = 100 Hz

b = 60 Hz

ts = 4 sec.

Q(0) = 150 Hz

P(0) = 200 Hz

n = 2

m = 60 Hz

then, rewriting **Q(0)** and **P(0)** as $\mathbf{Q_o}$ and $\mathbf{P_o}$ respectively for legibility, we solve for the equation:

(13) $\mathbf{t = [\log2(12*[sn*(P0 - r) + r]) - 12*\log2(Q_o/m)]/D}$

which is obtained by converting and setting **Q(t)** = **P(n)**. For these values we get **t = 2.6 sec**, a not unreasonable value. (We emphasize, however, the purely illustrative nature of this and other examples in the text. Among other factors, the slope **D** is language-specific, and English and Dutch do not necessarily share the same slopes.)

LP points out that peaks follow the same rule, no matter how far apart in time they are, and that the same difference should apply for the two peaks being 1 or 3 seconds apart. However, we can take Eq.(10) and easily (if tediously) solve it in terms of a given t, setting some other quantity as a variable: s, P_o, or using Eq.(2), ts. For example, this latter would mean that the longer time between the same peaks would indicate a longer sentence - not surprisingly.

If resetting forms part of the interpretation, introducing the variables with hats, then using at any given moment the adjusted formulae (4) and (14), one could solve for $\mathbf{P_o}$ ^ and **D** simultaneously, the interpretation being that the speaker resets his pitch and his state as he discovers how quickly or slowly he is speaking (accenting).

Resolution 6. Non-declining patterns.

In reading that downstep exists in the case of emphasis but not in the case of normal accenting, as LP asserts, one is tempted to raise the question as to the exact border between an accent and an emphasis. Since the border is assuredly not precise, and since a smaller pitch range will mean a smaller absolute decrease, they are very likely unrecognised downstep patterns. Nonetheless there are still cases of the upsteps and other non-declining patterns. We show that these fit into the integrated model, by dividing them up into three groups.

(Group I.) Upsteps can be the result of resetting mechanisms. Of these we distinguish two types: (a) a resetting of the pitch from one curve to the next; and (b) the same curve with regard to the frame of reference, but a resetting of the frame of frequency.

(Group II.) valleys, peaks, and flat sections on the line connecting the highs then are results of combinations of effects of Group I. To see this, and to show that this is perfectly in accordance with LP's formalism, we first recall our definition of utterance. In our model, a sequence of pitch accents X-Y-Z can yield an utterance of not only XYZ but also the utterances XY and YZ. If XY is an utterance of the "BA type" and YZ an utterance of the "AB" type, or vice-versa, then the end effect is a change by a factor of k*s (using the notation of section 2) which is approximately 1, giving the impression of an utterance XYZ with a v shape or inverted v, respectively, and an impression that XZ is flat.

(Group III.) As with all these analyses, it is possible that there are patterns which were handled by 'tH, and not by LP, and vice-versa. The most obvious differences are the language and cultural differences, but others may be of relevance.

Resolution 7: Final lowering.

On one side, we seem to be on solid ground with the concept of a lowering of frequency to mark the end of a declarative sentence: everyone agrees that this happens, and we even have the basis for a possible representation in a physiological frame of reference, in that there is a clear (physiological) release at the end of the utterance. (See the graphs of Collier, 1975.) That, therefore with respect to the horizontal axis of such a frame of reference, the curve need not show such a strong fall, if at all, would simplify the task of building the integrated model with a clear interpretation and make possible a simpler theory.

On the other side, when looking at the details, the ground seems to become shaky again. Is this phenomenon different from an extension of the pattern from the rest of the sentence ? LP says yes, 'tH says no. Is this a direct contradiction ? On the formal level, not at all, since the patterns in question are different. The difference may be a matter of formulation, but also perhaps due to a difference in domains. For example, since LP treated only English and 'tH only Dutch, there remains a question: is final lowering a language-specific phenomenon? If yes, then the rate for 1 will be around 0.7 for English -- i.e., for English L will be equal to 1 (small L),

and for Dutch, one may approximate also 1 (one). If no, then the question arises: does final lowering exist ? If no, then a bit of adjusting of the various constants in LP will do away with the need for 1 (or, of course, another formula). For example, if **s = 0.6, P = 200 Hz, r = 100Hz, and l = 0.7**, then one can calculate the F_o decay. One can then describe another decay with the same equations except without a final lowering constant, using **s = 0.54, r = 106 Hz.** In this case the decays are never more than about 2Hz away from one another! If, however, final lowering does exist, a very slight adjustment in 'tH's equations can also account for it. For this process we diverge for a moment to handle significant figures; this may seem pedantic, but it is important to explain our example.

Given any measurement or results of calculations based on measurements, one assumes that the numbers could have been rounded off; for example, given simply **"11"**, this could have been round off for any number between **10.5** and **11.5.** Likewise, a number **1.5** is really some number n, for **1.45 < n < 1.55.** So, the formula (2) above allows us to say that the slope for $\mathbf{t_s}$ **< 5 sec.** is given by

(14) **- 11.5/(ts + 1.45 sec.) < D < - 10.5/(t_s + 1.55 sec.)**

For example, if we have an utterance of exactly three seconds, the slope will vary between **-2.6 and -2.3 st/sec**. Using a base of **50Hz**, this can make the difference between the **148 Hz** and the **139 Hz** of the example for a final lowering of subject DWS in LP (compare Figure 21, p. 187 and Table 10, p.202, of Liberman and Pierrehumbert (1984)). Thus final lowering need not be contradictory to the declination of 'tH on the formal level, especially given the use of the 'uncertainty relation u' as explained previously.

There remains the contradictions on the interpretive level: what is the phonological phenomenon ? Does the final peak have a different decay or not ? From the preceding explanations, it is clear that the theories are compatible due to the application of the fuzzy function u, so that the integrated formal model could take either 'tH's or LP's interpretation and remain consistent (assuming it was consistent beforehand). However, u's

interpretation makes such a choice unnecessary: in denying the validity of such precision as appears in these other two models the integrated model differs equally from both; yet the presence of u assures that the predicted values in each are included in the range of predicted values in the integrated model.

7. The Conceptual Differences Again

By developing a formal mathematical model that integrates different properties of the two leading models of F_0 scaling, our purpose is not to overshadow the conceptual differences that underlie these models. We intend in this closing chapter to briefly present the major conceptual differences that need to be empirically tested. We shall also summarize some of the controversial points discussed above.

LP's model is based on level tones or peaks and eliminates the necessity of contour tones or tunes; the IPO model has pitch movements among its primary categories. The 'hat' model is for instance a combination of a rise and a fall. This is phonologically relevant, but may not be inconsistent, viewed from a quantitative perspective. LP's model does the same. The level approach integrates previous approaches that were configurational or contained contour tones. However, this is not to say that the IPO model is only phonetically oriented. As Ladd (1996:14) has stated it: 'The IPO tradition is in many ways the first to make a serious attempt to combine an abstract phonological level of descriptions with a detailed account of the phonetic realisation of the phonological elements."

LP's model doesn't reject an overall slope dependent on time. According to LP's empirical finding, even among variant speakers, the relationship between two accent peaks is constant, irrespective of the pitch range and of the length of the syllables. A local downstep, respective to a previous one, predicts the location of the next accent peak. Here again the choice between a downstep model and a time oriented model is based on phonology. Downstep is largely motivated by a slope independent on time among many languages, especially African languages. Time is a variable factor, and a continuous factor as opposed to tones that are discrete.

A fundamental conceptual difference with Fujisaki's model is that it is based on positive, High Peak tones (or accent commands) and doesn't take into account the Low (accented) tones. As pointed out by Ladd (1996:285) 'It is possible [emphasis, A&R] to approximate the low-rising contours, but this is inconsistent with the intend function of the phrase component'.

We would like to conclude this paper by quoting Ladd's (1996:285) fundamental remark that "(t)here seems little doubt that an overlay model is the best way to treat [microprosodic phenomena] in generating F_0 for synthetic speech'. If this is true, then we have shown how this type of model can be extended to formally integrate other different formal properties of F_0-scaling such as final lowering, lowering of successive peak accents and resetting.

8. Conclusion

When comparing two models, the present trend in linguistic F_0 scaling has focussed on the conceptual differences. However, the proponents of one linguistic model do not ignore the fact that even opposing models represent a considerable amount of data, as represented by the formulae as restricted to the conditions of the measurements (i.e., before extrapolation). Since any model should correspond to the tested and the testable, it is the theories which are primary which should form the basis for a unification. To convince the respective proponents and opponents of various theories that such a unification is even possible, one must show that the theories do not contradict one another on the formal level and need not contradict one another on the interpretative level. This is the raison d'être of the model presented here. Furthermore we have included in our formal model some elements which should invite the type of further experimentation so necessary to assure a proper correspondence between theory and interpretation.

ENDNOTES

3. We thank especially Dafydd Gibbon, Bob Ladd, Johan 'tHart and Rose Vondrasek who have made valuable inputs to this paper. The first author had the

opportunity to informally discuss a few of the issues with Mark Liberman who has been generous with his time. We are however responsible for all possible misinterpretations of the theories and all the errors contained in this paper. The first author would like to thank the Alexander von Humboldt Foundation for providing him with a grant that helped him to finalize the paper.

References

Ahoua, F. 1990. "Two current phonetic models of intonation analysis." *Cahiers Ivoiriens de Recherche Linguistiqu*e 25: 63-89.

Beckman, M. and Pierrehumbert, J. (1986). *"Intonation structure in Japanese". Phonology Yearbook* 3: 255-310.

Collier, R. (1975). *"Physiological correlates of intonation patterns." Journal of the Acoustical Society of America* 58: 249-255.

Cooper, W. and Sorenson, John (1977). "Fundamental frequencies contours at syntactic boundaries". *Journal of the Acoustical Society of America*. 683-692.

Fujisaki, H. (1981). "Dynamic characteristics of voice fundamental frequency in speech and singing." In STL-QPSR 1. 1-20; also in Peter F. MacNeilage (ed.) 1983. *The production of speech.* Heidelberg: Springer-Verlag, pp. 39-55.

Ladd, R. (1984). "Declination : a review and some hypotheses". *Phonology yearbook* 1: 53-74.

Ladd, R. (1990). "Metrical Representation of Pitch Register". *Papers in Laboratory* Phonology I. Ed. Kingston, J. and Beckman, M. Cambridge University Press.

Ladd, R. (1988). "Declination 'reset' and the hierarchical organization of utterances". *Journal of the Acoustical Society of America* 84: 530-44.

Liberman, M. and Pierrehumbert, J. (1984). Intonational Invariance under Changes in Pitch Range and Length. *In Language Sound and Structure*. Ed. Aronoff, M. and Richard Oehrle. Cambridge, USA: MIT,.

't Hart, J., Nooteboom, Vogten, L.L. and Willems, L.F. (1982). "Manipulation of Speech Sounds". *Phillips Technical Review* 40: 143-145.

't Hart, J. Cohen, A. (1973). "Intonation by Rule: A perceptual Quest." *Journal of Phonetics* 1.309-327.

Thorsen, N. (1983). "Two Issues in the *Prosody of Standard Danish". In Prosody*: Models and Measurements. Ed Cutler, A. and Ladd, R. Berlin, New York: Springer.

Umeda, N. (1982). "Fo declination' is situation dependent". *Journal of Phonetics* 10: 279-290.

Final Lowering
Time in Seconds

Appendix 2: Mark Liberman & Janet Pierrehumbert, " Intonational invariance under Changes in pitch range and length" (1984)

COMPONENTS OF THE INTONATIONAL MODEL

Q: What about Anna? Who did she come with?
A: Anna came with Manny.