INTONATION ANALYSIS: THE PERCEPTION OF SPEECH MELODY IN RELATION TO ACOUSTICS AND PRODUCTION

R. Collier
Institute for Perception Research
PO Box 513, 5600 MB Eindhoven, The Netherlands

1. INTRODUCTION

For approximately one century, the study of intonation has attracted a great deal of scholarly attention, despite the notorious difficulty to analyse by ear this evasive property of spoken language. Probably, the chief impetus to this long-standing effort stems from the conviction that speech melody has a particular communicative value. Indeed, individual speech sounds do not carry an intrinsic meaning of their own, whereas intonation and other prosodic features seemingly add something to the content of a message that is not already expressed in the semantics of its individual words, nor in their syntactic relations. Therefore, most intonation studies were to be found originally in traditional linguistics. However, even if intonation did not serve some communicative purpose, it would still be necessary to study in detail the melodic structure of utterances if one wants to design rules for the automatic control of fundamental frequency in the electronic synthesis of speech. Therefore, the study of intonation is also relevant for present-day speech technology.

In this paper I will discuss the nature and the advantages of a perceptual analysis of intonation. In particular, I hope to show that such an approach provides insights that are indispensable for the interpretation of data that are obtained in other domains of description, viz. speech acoustics and production. Also, an explicit melodic model is a prerequisite for the study of intonation from a functional point of view. I will present some of the essential properties of the perceptual study of intonation and in several discussion sections I will make suggestions for lines of research that could be pursued in order to clarify a number of questions evoked by this approach.

2. AN INTEGRATED APPROACH

It is convenient to deal with intonation at two levels of description: a concrete and an abstract one. At the concrete level, intonation manifests itself most directly in the form of perceivable speech melodies. Trained phoneticians can express such melodic impressions in the form of a transcription of a pitch contour, making use of notation symbols. At the same concrete level, intonation can be described in terms of those properties of the acoustic speech signal that give rise to perceived melodic variations, in particular: the changes of fundamental frequency (F0) over time. An equally concrete description of intonation can be provided in terms of the physiological variables that control the frequency of glottal fold vibration, for instance the activity of certain laryngeal muscles that affect the tension of the vocal folds. These three ways of looking at intonation are complementary phonetic approaches. A true understanding of the phenomenon at this concrete level implies that a description in one domain can be explicitly translated into a representation in the other two. The extent to which this goal can be reached is a measure for the amount of integration that we have achieved in our phonetic knowledge of intonation.

The infinitely diverse speech melodies that we perceive do not differ from each other to the same extent. Intuitively they appear to cluster into families of pitch contours, such that members of the same family are melodically more similar than those that belong to a different family. This perceived similarity can be accounted for by postulating that a family of similar contours derives from a common underlying intonation pattern. This assumption is analogous to positing a phoneme as the unifying concept for a set of audibly different, but phonologically similar speech sounds. Therefore, an intonation pattern is as abstract as a phoneme. An important goal in intonation analysis is to make explicit how concrete manifestations of speech melody relate to these abstract intonation patterns. By making this relationship clear we integrate our phonetic and linguistic insights of the phenomenon.

3. INTONATION FROM THE LISTENER’S POINT OF VIEW

The relation between the speech signal and the sound structure of language is a complex one. At the segmental level, speech perception can be considered as a transformation of a continuously varying acoustic spectrum into a sequence of discrete and invariant segments that correspond to phonemes. In the suprasegmental case of in-
intonation, the fundamental frequency of the signal is constantly changing, too. The analysis and interpretation of this parameter should also amount to a transformation that yields discrete and invariant descriptive units. It turns out that such units most readily and naturally reveal themselves in perception, namely in the form of pitch movements. Given the central position that the pitch movements hold in our descriptive framework of intonation, it is of great importance to be able to single them out in a reliable way.

3.1. Close copy stylizations

Traditional auditory analyses of intonation potentially suffer from the weakness that they are impressionistic and, therefore, unreliable. In order to give the perceptual study of pitch in speech a more solid methodological foundation, measures should be taken to verify whether the subjective impression that a given listener has, can be reproduced in others. Also, the objectivity of the perceptual approach can be enhanced if at all times the melodic impression can be related to an exact specification of the signal properties that give rise to it. These considerations have been at the basis of the "stylization method", developed by 't Hart and associates at the Institute for Perception Research ('t Hart and Cohen 1973; 't Hart and Collier 1975). In essence, the method allows one to determine which changes in Fo are relevant for the perception of speech melody. To this end, the capricious course of Fo is approximated with the minimum number of straight lines that are necessary to imitate the original utterance without any audible difference. By definition, the straight lines correspond to the perceptually relevant Fo variations, while all the physical changes that can be smoothed out are irrelevant for perception of speech melody. The resulting pitch contour is called a "close copy" (De Pijper 1983). Although it is a perceptual entity, it remains firmly anchored in physical reality: the constituting pitch movements all have a definition in the domain of (logarithmic) frequency over time. Straight line approximations do not sound different from exponential ones, but have the advantage that the pitch movements can be represented literally as discrete units: they have clear beginning and end points in a two-dimensional space.

It turns out that pitch movements differ in more than one respect: direction (rising, falling,...), rate of change (abrupt, gradual,...), size (large, small,...) and position in the syllable (early, late,...). The number of distinctions that are necessary along each dimension can be assessed in listening experiments and the attributes of all the distinct pitch movements can be defined in physically interpretable terms (semitones per second, milliseconds after vowel onset, coincidence with end of voicing, and the like). Figure 1 illustrates the physical difference between an original Fo curve (a) and its perceptually iden-
tical "close copy" (b). Whether such close copies are indeed perceptually indistinguishable from their originals can be verified in formal listening tests.

3.2. Discussion

The IPO investigations of Dutch, English, German and Russian intonation have confirmed the usefulness of the stylization technique. Yet, the "close copies" are still surrounded by many questions. For instance, the straight lines smooth out all the Fo perturbations in vowels, caused by consonants (the so-called "microintonation"). In syllables without a pitch accent they also do away with the effects of the intrinsic pitch of the vowels. How can it be explained that these Fo changes, which are sometimes of considerable magnitude, are not perceived? Another question concerns the uniqueness and reproducibility of a "close copy". How much correspondence is there between close copies of the same utterance, made by different experienced researchers? What is the nature of the differences that are found? Are such differences audible? If not, what are the perceptual tolerance regions within which stylized contours do not audibly differ from original Fo curves? Can one really make a close copy of an utterance in a language that one does not understand? In other words: does (implicit) knowledge of the intonational structure of the language under study affect the outcome of the stylization procedure? Such questions have a bearing on the foundations of a method that has been very successful in practice, but still needs a more solid theoretical underpinning, for instance in psycho-acoustic terms. At IPO work is in progress that attempts to provide such a basic insight into the nature and the process of Fo stylization.

3.3. Standardization

The various pitch movements that occur in close copies may differ from each other in many ways. Yet, in order to achieve a certain degree of generalization, necessary for the development of speech synthesis rules, the different types of movement should be given standard specifications for their relevant features. Thus, for instance, a typical accent-lending rise in Dutch has a duration of 120 ms, its excursion size is 6 semitones and it is timed in such a way that its peak lies 50 ms after the vowel onset of the accented syllable. Standardization also concerns the permissible sequences of pitch movements. A correct pitch contour consists of a lawful string of rises and falls. Most of the theoretically possible combinations of pitch movements do not occur, some sequences are recursive while others are not, etc. These constraints can be formally expressed in a so-called "grammar of intonation", which may take the form of a transition network. Each possible trajectory through the network defines an acceptable contour. Of
course, such a grammar is the core of a scheme to be used in the automatic generation of artificial pitch contours (Willems, Collier and 't Hart 1988). Figure 1 (c) shows a standardized contour as generated in synthetic speech.

**3.4. Discussion**

For elementary intonation synthesis it may be sufficient to have just one standard recipe for each distinct type of pitch movement (but see Collier and Terken 1987). In natural pitch contours, however, there is a much greater variability so that in intonation perception the listener has to decide which change of Fo corresponds to which pitch movement category. Presumably he bases this decision on tacit knowledge of the relevant signal properties, but it is very likely that he supplements this "bottom up" approach with a "top down" strategy, using knowledge of the intonational structure of the language in question. One way to model how a listener may proceed, is to simulate the assignment of Fo changes to pitch movement categories in an algorithm for automatic pitch transcription (Boves, ten Have and Vierregge 1984). At IPO we have developed software that automatically stylizes Fo curves into pitch contours (Hermes and 't Hart 1987). Ideally, such simplified contours should be equivalent to "close copies" as made by experts. We are planning research in which the output of this algorithm would serve as the input to a system that automatically transcribes the pitch movements in terms of the 10 relevant categories that we distinguish in Dutch ('t Hart and Collier 1975). It will be interesting to see how much information can be extracted from the signal and how much additional knowledge has to be supplemented concerning higher levels of intonational structure. As far as information contained in the signal is concerned, a relevant categorization cue is the timing of pitch movements with regard to vowel onsets. The system will therefore have access to the output of an algorithm that detects vowel onsets (Hermes 1988). Information of a higher order could be provided in the form of an explicit statement of the sequential constraints of pitch movements, as formalized in a "grammar of intonation".
4. A PERCEPTUAL DETOUR

The choice of a perceptual vantage point in the analysis of intonation is a debatable decision. If one is interested in human communication through spoken language, the perceptual approach would seem a natural choice since, after all, the listener can only use his ears to receive the message. But the rather cumbersome methodology needed to counteract the inherent subjectivity of the perceptual approach makes it less attractive than a more direct and objective measurement of the acoustic or physiological features of intonation. The perceptual analysis is a detour, and the question is: does it pay?

4.1. In relation to acoustics

In earlier days there was a perfectly practical reason to rely on auditory analysis, since there were no reliable pitch meters. Now that this technical difficulty has been surmounted and numerous well performing pitch measuring devices are available (Hess 1983), we are faced with a new problem, namely that such meters reveal a wealth of detail that is extremely hard to interpret in terms other than frequency over time. In particular, the minute variations in F0 recordings do not stand in any straightforward and transparent relation to the perceptual impression that one has when listening to the utterance that was measured. In Figure 1 one can see how large the discrepancy between physical and perceptual reality can be. Therefore, the advocated perceptual approach amounts to a form of data reduction that singles out those physical changes that are important for the listener. Notice, however, that the end product of the perceptual analysis is more than an auditory impression; it is also an exact and explicit specification of the signal properties that give rise to that impression. In this way the perceptual detour does not replace the acoustic measurement; it makes it interpretable in terms that are potentially relevant from a communicative point of view. In this respect, the detour is worthwhile.

4.2. In relation to production

Also in the other domain in which a more direct measurement of intonation is possible, namely speech physiology, the perceptual approach has its merits. The production measurements related to pitch control in speech are mostly of two kinds: electromyographical recordings of the activity of selected laryngeal muscles and registrations of the subglottal air pressure variations. Such signals are known to be very noisy and even after processing (filtering, rectification, smoothing) they remain difficult to interpret. Figure 2 shows an electromyographic recording of the vocalis muscle during the production of the Dutch utterance "Heleen wil die kleren meenemen" (Helen wants to take these clothes along). This muscle corresponds to the internal portion of the thyro-arytenoid muscle, i.e. the vibrating part of the vocal folds. Contraction of this muscle increases the tension of the folds and makes them vibrate faster. Consequently, a burst of activity is expected whenever the speaker wants to produce a perceptually relevant rise. Therefore, knowledge of the melodic structure of the utterance under analysis will make the interpretation of the physiological data easier, especially if the pitch contour is complex.

This potential merit of the perceptual detour can be made plausible on the strength of the following assumption: perceptually relevant pitch changes are those that have been produced by some voluntary action on the speaker's side. Having tacit knowledge of the selective attention that a listener pays to F0 variations in speech, the speaker will actively control the realisation of precisely those variations that contain communicatively important cues. Therefore, it is likely that the most relevant physiological information will be found in the vicinity of perceptually important pitch changes, caused by discrete commands to the laryngeal and/or respiratory muscles.

4.3. Discussion

In speech recognition the listener attempts to establish a link between the continuously changing spectral properties of the acoustic signal and a linguistic representation of the utterance in terms of meaningful units, such as words. In the prosodic domain he performs a comparable operation, viz. he interprets the highly variable course of F0 as a coherent melodic structure, an intonation pattern. In both cases the interpretative process can be modelled by postulating a level of structure at which discrete units play a role: speech sounds and pitch movements, respectively. Whether such units actually constitute intermediate levels of awareness during the recognition process or are analytical elements that become accessible only after recognition of the whole structure has taken place, is an interesting question. For intonation perception it is far from clear what role the individual movements play, but there are indications that they do not merely blend into one "global" percept, but rather (also) stand out as "atomistic" elements. For instance, pitch accents are local phenomena, often caused by the presence of a single rise or fall. The same holds for the isolated rise or fall that marks the boundary between syntactic constituents. Therefore, playing such a central role in perception, pitch movements can guide the investigator's attention when he analyses and interprets acoustic or physiological data. Advance knowledge of the great wealth of melodic possibilities in a language can also inspire the researcher who is interested in the physics or the physiology of intonation, to select a more diversified set of contours than the usual sequence.
Figure 2. Fo curve (a) and the electromyographical recording of vocalis muscle activity: high pass filtered (b), integrated (c) and smoothed (d). The utterance is "Heleen wil die kleren meenemen" (Helen wants to take these clothes along). [Courtesy of H. Strik, Univ. of Nijmegen, The Netherlands].

of rise-fall combinations (Collier 1975).

In summary, the perceptual detour pays off in a double respect. It brings into focus the perceptually relevant pitch movements which, as natural descriptive units, reveal the melodic organisation of an otherwise unstructured Fo curve or physiological registration. Secondly, the firm and explicit link between perception and acoustics provides the basis for the rule-governed synthesis of a highly variegated set of pitch contours.

5. DECLINATION

The most important global attribute of an Fo curve is the tendency of the frequency to decrease slowly from the beginning to the end of an utterance. This phenomenon was named "declination" by Cohen and 't Hart (1965).

5.1. Perceptual relevance

A first indication that declination is a perceptually relevant attribute of pitch contours is that synthetic intonation without declination does not sound natural. So how can declination be modeled? There are several alternatives. It could be that the effect arises from the fact that, in general, falls are larger than rises. It turns out that artificial pitch contours that simulate this property do not sound acceptable. Such a recipe implies that, in between rise-falls, Fo minima remain strictly monotonous. This produces a highly unnatural impression. A second alternative is to assume that declination is a property of Fo maxima, rather than minima: declination is a "topline" through Fo peaks (Cooper and Sorenson 1981). The drawback of this model is, that the height of Fo peaks correlates with the strength of the pitch accents, so that it becomes impossible to fit a line through the maxima if the prominence relationships between successive words are highly varied. Also, topline declination cannot be established if there is only one Fo peak in an utterance. A third way of interpreting declination is to consider it as a local rather than a global phenomenon: it is not an overall property of the Fo curve, but a scale factor applied to each Fo peak individually. This "down step" idea is advocated by Liberman and Pierrehumbert (1984). Again, the actual height of Fo peaks is not a strictly intonational property; to correctly model it, requires knowledge of the degree of emphasis a word has to receive. For the languages that we have studied, namely Dutch, British English, German and Russian, the most
5.2. Measuring declination

One way to measure declination acoustically is to fit a straight line through the first and last data points in an Fo curve. This will often cause problems, since the presence of an extra low value before the first pitch rise and/or after the last fall will make it impossible to fit a straight line through most of the valleys. For this and other reasons the visual fit of an hypothetical declination line to actual Fo data remains disputable technique. A better way is the use of a "close copy" stylization. If the lower stretches of pitch can be replaced by pieces of one straight line, without perceptual consequences, it can safely be assumed that the "actual" declination line will be indistinguishable from the measured one.

5.3. Communicative value

It has been observed that the slope of the declination line varies with the length of an utterance. However, in fairly long utterances the "close copy" technique reveals that the pitch contour is broken up into two or more, shorter prosodic phrases. In such cases there often appear so-called declination "resets" at the prosodic boundaries (Cooper and Sorensen 1977). The occurrence of such new starts of the declination ramp has already been studied from a physiological and an acoustic point of view (Collier 1987), but many interesting functional questions still wait for an answer. In particular, if such resets are perceptually conspicuous, they are potential cues for the listener with regard to the syntactic organisation of the utterance. Indeed, there are indications that resets preferably occur at syntactic boundaries and that their size may correlate with the depth of those junctures (Appels 1985, Ladd 1988).

5.4. Discussion

The use of "close copy" stylization is only a tool to establish the course of the declination line perceptually. We do not really know yet which attributes of the Fo curve contribute to the perceived course of the baseline. For instance, it is unclear how this course is affected by initial or final syllables that are much lower than the intermediate valleys. Also, it is still unknown how such valleys influence the overall perceptual estimation if they do not fit on one straight line. The answer to such questions will eventually lead to algorithms that can automatically determine the perceptual trajectory of the declination line on the basis of measured Fo values.

At IPO, work in this direction has been planned for the near future. In the communicative domain, too, important issues are still unresolved. Most of our knowledge about the course of declination is limited to properties of isolated, read-aloud utterances. It has become increasingly evident that existing models of declination cannot cope with the more complex reality of coherent texts. For instance, in cases of declination "reset" it is still has to be studied in detail what the onset and offset frequencies should be of the non-terminal and terminal baseline stretches. Also, it needs to be examined whether a fundamental distinction should be made between "resetting" of the baseline and an actual upward "register shift". The latter may be an appropriate marking of a new information unit, such as a paragraph, whereas "resets" are bound to occur within one utterance.

6. CONCLUSION

Although the analysis of intonation can in principal start in any domain of phonetic description, be it physiological, acoustical or perceptual, it appears to be profitable to choose the perceptual vantage point. Indeed, the perceptual approach acts like a filter that singles out those aspects of the signal that are of potential relevance for speech communication. Therefore, a working perceptual model is also a prerequisite to ask well guided questions about intonational function. Finally, if the model is firmly anchored in the acoustic domain, it provides an excellent starting point for the development of an intonation-by-rule system in synthetic speech applications.

REFERENCES


