Proceedings of the Institute of Acoustics

A TONAL ATTRACTOR MODEL OF ENGLISH INTONATIONAL PHONOLOGY, AND ITS IMPLICATIONS FOR F0 IN SYNTHETIC SPEECH.

T. Gillott
BT Laboratories
Martlesham Heath
Ipswich IP5 7RE

1.0 INTRODUCTION

This paper describes a phonological model of intonation developed at BT Labs and which is currently undergoing tests in order to assess its suitability for inclusion in "Laureate", the BT Text to Speech system. It is in many ways a descendant of the Tonal Sequence intonation model which was developed by Janet Pierrehumbert [1]. The current Laureate intonation system is based on a version of the model described by Silverman [2].

After some years of using the Silverman/Pierrehumbert system, it became apparent that, excellent though it was in many ways, it had a number of limitations. These limitations were at least in part a consequence of the theory behind the model, and thus not easy to change.

It was felt that a modification of parts of the model might produce useful improvements, but in the event a new theory was developed which uses the most effective parts of the old model and, hopefully, circumvents the problems produced by the less effective parts.

The theoretical stance adopted in this paper is that the Tonal Sequence (TS) concept is a sound one for modelling F0 in synthetic speech. That is, the notion of a series of discrete events in the phonological domain forming the basis of a continuous stream of acoustic values is valuable for the following reasons:

(1) It separates the phonological component of intonation from the phonetic one, unlike many other models

(2) This separation allows new sets of phonological driving rules to be developed without any reference to the actual frequency values used

(3) Values can be readily inserted from new voices without having to perform a re-analysis of the data in some tune inventory

All these concepts are faithfully adhered to in the new model. Further, it will be shown that the new model improves upon conventional TS models in a number of ways.

Firstly, its phonological inventory is simpler. In particular, the concepts used allow the inventory to dispense completely with the Phrase Accent component with no loss of descriptive power. As well as allowing many inconsistencies in the TS theory to be removed, this increased simplicity eases implementation considerably.

Secondly, the concept of phonological separateness is shown to be adhered to more strictly under the new model. It is argued in this paper that many of the properties of the phonological part of the Pierrehumbert model are actually phonetic in nature, and thus should be part of the acoustic-phonetic realisation component.

Thirdly, it is claimed that this model is a more natural one than conventional TS systems in that it assumes maximal simplicity in the psychological-phonological domain, mapping to maximal complexity in the acoustic-phonetic domain. This reflects the idea that a relatively

simple command produces quite complex effects, without having to specify the mechanics of these effects at the command level. This falls more in line with current thinking on cognitive processes [3].

The paper firstly introduces the concepts on which the model is based. It goes on to propose a re-analysis of the phonological tone inventory, showing how the concepts just introduced allow this. This re-analysis also reveals the phonetic nature of several of the components of a conventional TS model that were previously presented as being phonological in nature. Finally, a few of the advantages to a Text to Speech system of using this new model are described, with particular reference to the BT Laureate system.

2.0 THE TONAL ATTRACTOR THEORY - BASIC CONCEPTS

TS models in general postulate a series of events which form the phonological basis for an acoustic-phonetic event. These events have associated with them a form of phonological "level", usually referred to as prominence, salience or something similar. This prominence is then mapped at the phonetic level onto frequency values in Hz of the fundamental contour.

In the Pierrehumbert TS model, these *pitch events* have both shape and duration. These parameters control the shape of the F0 contour at each important point, in effect both forcing F0 into certain paths and forming a phonological "scaffolding" on which the eventual contour is built.

This idea, although basically a sound one, has a number of problems when it comes to implementation. A number of ways of implementing the model were tried under the BT system, but it gradually became clear that the theory as expressed in [2] had some limitations in certain areas, notably in the behaviour of F0 in short fall-rises.

In the following description it is assumed that the reader is familiar with the basic concepts of a TS model, but the notion of reference line is important in the following discussion, so it will be outlined here.

A speaker has a natural range within which his or her F0 operates. The top of this range is often referred to as the ceiling, and the bottom as the floor. One further value is termed the reference line. This is a pitch value where the speaker feels most comfortable, and is used as reference point from which pitch excursions depart and to which they return. Because this pitch value remains fairly constant throughout an utterance, it is envisaged as a line, hence the term.

It was decided to express phonological pitch events using a different method from those in conventional TS models, one which would overcome the shortcomings described above when realised at the phonetic level. This method is underpinned by the view that F0 may be perceived as the phonetic-acoustic equivalent of a series of pitch movements away from the speaker's reference line. The forces that move the pitch are placed at salient points of the utterance. Borrowing from the terminology of dynamical systems theory [4], these forces are referred to as *attractors*. Because they act on pitch or tone, they are qualified as *tonal attractors*.

A tonal attractor operates in the perceptual space delineated by the floor and ceiling of the

speaker's range and the time taken for an utterance to take place (the so-called T (Tone) space). The reference line has a constant value in relation to the range extrema, and is seen to extend over the entire time of the utterance. If left undisturbed, pitch would describe a flat line which followed the path through the T-space described by the reference line.

A tonal attractor produces a strong attractive force at a point in the T-space, and pitch will follow the path from the reference line dictated by this force. It will then return to the reference line.

There is a remarkable lack of restrictions on tonal attractors. The only real requirements are that they should exert an attractive force either up or down from the reference line, and that they should occur at a specified time in the utterance. Thus it is possible to see that a very simple inventory may be built up from the concepts of attraction. This will be dealt with in more detail in 3.0 below.

This notion of attraction to a point in the T-space away from the reference line is not in itself sufficient to allow a full description of the behaviour of pitch. In the absence of any attracting accents, pitch will tend to move back towards a reference point which differs from speaker to speaker. It is thus logical to postulate that the reference line itself functions as an attractor in the T-space, not merely as a point of reference as it does in most TS models.

The amount of influence exerted by an accent varies and is directly dependent on its prominence. The concept of prominence remains the same in the tonal attractor model as in conventional TS models. Influence is not confined to the upward and downward directions; an accent is deemed to have influence in the time domain as well. It is the interaction between these conflicting attractions which forms the basis for pitch movement, and thus eventually for an F0 contour.

3.0 A PROPOSAL FOR A SIMPLIFIED PHONOLOGICAL INVENTORY

It was mentioned in 2.0 above that the concepts of attraction by both the pitch events and the reference line of the speaker's F0 grid allow a simple inventory of phonological units to be built up. This assertion will now be amplified and examined in some detail.

In the Pierrehumbert-type TS model, pitch events are subdivided into various types. There are pitch accents, which indicate where a pitch excursion is to occur in a stressed vowel. Further to these, there are phrase boundaries, which indicate where pitch is to end up at the end of a phrase, and phrase accents, which are a special case in a number of ways.

It is asserted here that such a diverse inventory is not necessary. In particular, the events controlling pitch movement after pitch accents and at the ends of phrases should not be distinguished from those which are designated as pitch accents.

In this inventory, there are only two types of attractor: those which attract pitch upwards from the line, specified as High (H), and those which attract it downwrds from the reference line (Low or L).

At first sight, this would seem to be inadequate to describe the relatively quick and complex F0 movements observed in events such as fall-rises. In a Pierrehumbert-type system, these

are modelled with a sequence such as high pitch accent, followed immediately by low phrase accent, followed by high phrase boundary. It will be recalled, however, that in the Tonal Attractor model the reference line itself acts as an attractor. Thus, if two H-attractors are placed at intervals during the utterance, the pitch will first be attracted to the first H-attractor as we move through the utterance. However, as we move away from the first H-attractor, its influence diminishes, and the pitch value comes more under the influence of the reference line than either of the two H-attractors. When we have moved further towards the second H-attractor, it in turn begins to exert more influence, thus overcoming the influence of the reference line attractor. The resulting pitch, if plotted, will be seen to describe a fall-rise pattern.

The motivation for getting rid of the phrase accent is that of rationalising the model. The phrase accent is a largely artificial category, which is only in the inventory to model a special subset of pitch behaviour. It behaves exceptionally in a number of not very well-motivated ways. This may be seen in the implementation produced by Silverman, in which many rules in the sets designed to ensure correct alignment of pitch events in the utterance contain exceptions geared specifically to deal with the phrase accent. Silverman himself [2] comments on the irrational behaviour introduced into the model by this event category, making particular reference to the way they require special treatment in phonological and phonetic lookahead. Ladd [5] criticises the fundamental need for phrase accents, and proposes a re-analysis of intonation without them.

If we assume, following Ladd, that the phrase accent is a redundant category, then we must have some alternative method of describing the behaviour it purports to model. The model for a fall-rise described above - two H-attractors only - leaves us with a problem. How can we distinguish between the temporary "sag" in F0 between accented areas, and the much quicker movement observed at nuclear accent patterns, using just two types of attractor?

To answer this question it is necessary to examine just why phrase accents were seen as being necessary in the first place. They are associated with nuclear pitch accents only, and they are there to force quick pitch movement in these positions.This is in accord with the observed phenomenon of quickly moving F0 at the most salient (ie nuclear) part of an utterance. Quick F0 movement produces an emphatic effect, and serves as part of the acoustic-phonetic signal to a listener that this part of the utterance should be given special attention.

Straight away we have a single underlying principle whose acoustic realisation can be observed in natural speech. But why describe it phonologically by inventing a new category in the inventory? It is argued here that a slight modification of the behaviour of the nuclear attractor is all that is needed. These already shift their position in time in a conventional TS model in order to indicate nuclearity.

The behaviour modification that is suggested is that the influence exerted by the nuclear attractor is single-sided only. That is, instead of exerting influence equally both before and after the point at which it occurs, a nuclear attractor exerts influence backwards only. It exerts none forwards.

The effect of this is to remove any pull towards the attractor once the pitch in the utterance occurs later than the attractor in time. The forces acting on it are a stronger than normal pull

towards the reference line, as are subsequent influence exerted by the following H-attractor.

The stronger than normal pull towards the reference line will result in a much quicker than normal movement of pitch. In fact, it is argued that the sudden lack of influence will actually cause the pitch to overshoot the reference line and plunge towards the floor of the grid before the direction of its movement can be reversed by the combined influence of the reference line and the next H-attractor. This accounts for the wider than normal range of F0 movement observed at the phonetic level in these positions.

Thus by slightly modifying the behaviour of an existing category we are able to describe a whole range of observed phenomena. Even bitonal accents (whose validity in Standard British English is highly dubious anyway) may be described in these terms.

As well as reducing the number of elements in the inventory, it is possible to reduce the complexity of the nature of the pitch events themselves. In a conventional Tonal Sequence model, pitch events have a complex set of characteristics associated with them, many of which are claimed to be phonological in nature.

Among these is the underlying shape specified for various events, and in particular pitch accents. This is expressed in terms of "level" and duration. Pitch events are described as having a shape similar to that of an inverted "L". They must be described at every point, starting with the bottom of the "leg" of the L, and carrying on to the flat part, etc. Many of the complex rules specified for pitch accent overlap deal exclusively with trimming these parts into various new shapes depending upon whether another accent is too close for comfort.

While this is a perfectly possible way of describing an accent, it has a number of weaknesses. Firstly, there is the complexity of the rules needed to ensure that the F0 resulting from the placing of accents behaves in a perceptually correct manner. This shows up most when attempting to model Fall-Rises in short syllables. It is very difficult to get all the overlap and trimming rules to work properly in this case without pitch event overlap making it difficult for a convincing F0 contour to ber realised at the phonetic level.. A number of potential solutions have been tried in the BT system over the years, but none have been fully satisfactory: Fall-Rises have remained the weakest part of the model. It is submitted that the over-complex nature of both the presumed underlying phonological shape of the accent and of the subsequent rules needed to control it is the root cause of this difficulty.

Secondly, it is doubtful that the phonological nature of an event should have to be described in such detail. All detail about shape should be confined to the acoustic realisation of the phonological component, thus ensuring maximum simplicity in the phonological area. It is further suggested that any such description that purports to be phonological in nature is really phonetic. Thus, the phonological and phonetic components of the conventional Tonal Sequence event inventory are by no means as well separated as they should be.

4.0 ADVANTAGES OF A TONAL ATTRACTOR MODEL TO A TEXT TO SPEECH SYSTEM

A Text to Speech system is very complex. Any theory which may lead to a lessening of this complexity is to be desired. One of the main features of the tonal attractor model is its ex

treme simplicity coupled with its powerful nature. Most, if not all, of the commonly occurring intonational phenomena in English can be described, and yet the rule set used to do this is much smaller than that used by a conventional Tonal Sequence model.

One reason for this is the use of the reference line as an opposing force to the accents. It obviates the need for one entire class of event, and thus also the set of rules used to describe this event class behaviour. This rule set is also the most complex and ad-hoc one, which reflects the assertion made earlier that the phrase accent was probably the least well-motivated of all event classes in the previous model.

Another reason is that this model produces complex effects by means of extremely simple commands. The idea of a single High or Low accent primitive is a simple one, yet when the attractive forces between accents and reference line have been balanced at the phonetic realisation stage, some complex F0 movements may be observed as a result. This is desirable in the light of modern research into cognitive phenomena.

5.0 CONCLUSION

This paper has presented a brief introduction to the theory of Tonal Attractors as applied to pitch in synthetic speech. It has been seen that it offers advantages over conventional Tonal Sequence models, from which it is descended, in terms of theoretical simplicity. It is hoped that applying this theory to the "Laureate" system advantage may be gained both in computational efficiency and in naturalness of speech.

6.0 REFERENCES

[1] Pierrehumbert, J. B. The phonology and phonetics of English intonation. PhD Thesis, MIT, 1980.

[2] Silverman, K.E.A. The structure and processing of fundamental frequency contours. PhD Thesis, University of Cambridge, 1987.

[3] Liberman, A.M., Cooper, F.S., Harris, K.S. and McNeilage, P.S. A Motor Theory of Speech Perception. Speech Communication Seminar, Stockholm 1963

[4] Tishby, N. A dynamical systems approach to speech processing", Proc. IEEE ICASSP '90, Albuquerque, Vol. 1, pp. 365-368, April 1990.

[5] Ladd, D. R. Phonological features of intonation peaks. Language, 59, pp. 721 - 759, 1983.