K-Max: a tool for estimating, analysing, and evaluating tonal targets

Antoin Eoin Rodgers1

1 Phonetics and Speech Laboratory, Trinity College, Dublin
roderan@tcd.ie

Abstract

This paper presents a novel approach to the identification of tonal targets within the Autosegmental Metrical (AM) framework using the second time derivative of the \( f_0 \) contour. The approach is implemented through an interactive Praat script called K-Max, which allows users to annotate salient turning points on a text grid as well as correct tracking errors and remove micro-prosodic events on the pitch contour. The script also generates a resynthesized model of the pitch contour based on the annotation of turning points, which is not typical in the AM approach. The theoretical rationale for the overall approach is presented, followed by a description of its implementation. The paper then discusses the success of the technique in identifying tonal targets in relation to user intuitions and acceptability judgments regarding the resynthesized \( f_0 \) contours. Finally, it provides examples of its potential application regarding the apparent duration of tonal targets, such as via tonal spread 교수

1. Introduction

1.1. AM and identification of tonal targets

In the Autosegmental Metrical approach (AM) to intonation [1], the fundamental frequency (\( f_0 \)) contour is viewed as being, in part, the result of the phonetic implementation of a string of underlying low and high tonal primitives (L and H) which are independent of but associated with landmarks in the segmental string and metrical structure. These tonal primitives can form pitch accents, which can be monotonal (H*, L*) or bi-tonal (L*+H, H*+L, etc.), and occasionally tri-tonal. The starred tone indicates the tone associated with a stressed syllable in the metrical structure. Tones can also manifest as edge tones, such as boundary tones and phrase accents. While the tonal sequence is considered highly abstract [2], at the same time, however, much research has been conducted to ascertain how phonological tones are linked to or aligned with elements in the metrical structure and segmental string [3]. This involves precise measurement of the timing of L and H targets, typically in pitch accents, in relation to landmarks such as syllable onsets or the onset of vowels in stressed syllables. In other words, manifestations of these highly abstract primitives can still be identified and measured empirically. As such, tonal targets are identified in the \( f_0 \) contour in terms of turning points. These can be either \( f_0 \) maxima and minima, or elbows in the contour, the latter referring to points where there is a distinct shift in \( f_0 \) trajectory.

As implied in the use of the terms High and Low, \( f_0 \) maxima and minima are the archetypal turning points, and elbows their less ideal manifestations. \( f_0 \) maxima and minima are also easier to identify than that of elbows. Within the AM approach, one common means of estimating elbows without simply eyeballing the curve is via a technique which involves line fitting two lines inside a designated portion of the contour [4]–[6]. The intersection of the best fit lines is taken to indicate the timing of the elbow. An alternative means of measuring turning points includes using the extrema of the second time derivative of a smoothed \( f_0 \) contour, or \( f_0''(t) \) [7].

1.2. K-Max: aims and guiding principles

K-Max [8] is a tool developed using Praat [9] scripts. It was developed based on the AM assumption that intonation can be described as a sequence of H and L targets. However, rather than viewing \( f_0 \) extrema as prototypical turning points, it takes the view that \( f_0 \) extrema just happen to be the most salient form of turning point. For this reason, as will be discussed in section 2.1 below, turning points are estimated using \( f_0''(t) \).

K-Max is designed primarily to facilitate the identification and analysis of tonal targets in pitch accents. However, it is also designed for the analysis of phenomena which are potentially more problematic within the AM approach, most notably, \( f_0 \) plateaux and valleys. The difficulty with plateaux and valleys is that they have duration, and as such are not readily identifiable as tonal targets. AM approaches have tried to account for the apparent duration of tonal targets, such as via tonal spreading in, in which one tonal target—typically a trailing tone—is seen to extend beyond the initial target time [10]. Different techniques have been used to quantify plateaux, such as [11], in which the plateau edge is measured heuristically in terms of a percentage fall from \( f_0 \) peak. Since K-Max provides a method of identifying and quantifying turning points in a principled manner, not only for pitch accents, but also for plateaux and valleys wherever they appear salient, it can also help provide empirical data in the analysis of features such as tone spreading.

A further aim in developing K-Max was to permit analysis by resynthesis using turning points identified as salient by the analyst. In part, this stems from the view that empirical analysis of tonal targets should contribute towards the effective and efficient modelling of the pitch contour for (re)synthesis, a view not necessarily widely held within AM [2]. More importantly, this goal is based on the view that it is important to demonstrate that a contour can be modelled effectively using theoretical principles of the AM approach to help demonstrate its validity. It is believed that in modelling the contour more comprehensively, elements of phonetic implementation which might otherwise be over-looked will have to be accounted for.
2. Quantifying tonal targets and slopes

2.1. Turning points and the second derivative of \( f_0 \)

If we consider \( f_0 \) or the rate of vibration of the vocal folds in terms of (angular) velocity \([12]\), its first derivative is the rate of change of velocity—i.e., acceleration—while the second derivative, \( f_0''(t) \), is the rate of change of acceleration, also described as jerk. The intuition that jerk is salient can be understood by analogy: in term of motion, jerk is experienced as the sensation one gets in a car as it begins accelerates, while peak jerk occurs when a car breaks suddenly.

\( f_0''(t) \) extrema align temporally with points of maximum curvature (concavity or convexity) in the \( f_0 \) contour. It should be noted that while \( f_0''(t) \) is not a direct measure of curvature, for convenience, these time points of the extrema will be referred to as \( K_{max} \). In the \( f_0 \) contour, these points are evident as turning points at \( f_0 \) extrema and elbows. The nature of this relationship between \( f_0''(t) \) and turning points is readily apparent if we consider the linear stylised contour shown in Figure 1. \( f_0''(t) \), shown in red, spikes at turning points in the \( f_0 \) contour (black), regardless of whether they are elbows or \( f_0 \) extrema. Negative spikes coincide with the most convex points—often \( f_0 \) maxima—and positive spikes with the most concave points—typically \( f_0 \) minima. Thus, both \( f_0 \) elbows and extrema are manifestations of the same phenomenon, and the polarity of \( f_0''(t) \) indicates whether the turning point is more H-like (negative) or L-like (positive) at that time point. It is worth noting that in real-world cases \( f_0 \) extrema often occur near rather than at \( K_{max} \), so \( f_0 \) maxima and minima may even sometimes be viewed as symptoms of or epiphenomena around \( K_{max} \).

![Figure 1: Stylised \( f_0 \) contour and its second derivative.](image)

2.2. Tonal targets, physiological constraints, and planning

In order to distinguish tonal targets as described in the PENTA model from those of the AM approach, Xu describes articulatory targets in the PENTA model as covert as opposed to the overt tonal targets of the AM approach \([13]\). Overt targets, on this view, are measured in terms of literal \( f_0 \) maxima and minima as realised in the contour. Covert targets, on the other hand, are underlying articulatory targets which are not realised literally in the \( f_0 \) contour. This is because physiological constraints on the larynx limit the speed at which pitch trajectories can change \([14]\), resulting in phenomena such as tonal undershoot which occurs as speakers must adjust their \( f_0 \) trajectory over time \([15]\).

The argument that tonal targets should not be viewed as co-equal with surface manifestations of \( f_0 \) is compelling. However, it does not seem incompatible with the AM approach. While it has indeed been that case that AM analyses of tonal targets have measured tonal alignment and scaling in terms of literal targets, such as in \([16]–[18]\) inter alia, this may sometimes be more a case of methodological convenience than anything else. In fact, within AM literature, the \( f_0 \) contour itself is viewed as the result of the phonetic implementation of a phonological surface representation \([19]\). As such, it is reasonable to argue that adjustments made to \( f_0 \) during phonetic implementation as a result of physiological constraints are responsible for differences between phonological surface representation and its realisation in the \( f_0 \) contour. Thus, within the AM approach, one might well expect a mismatch between literal targets in the acoustic signal and the ideal targets of the phonological surface representation.

2.3. Trajectories and Inflection points

One potential strategy then is to estimate the ideal trajectory of the \( f_0 \) contour towards the ideal tone minus the effects of physiological constraints. Bearing in mind that \( f_0''(t) \) minima and maxima indicate points of maximum convexity and concavity respectively in the \( f_0 \) contour, the points of zero curvature between these two points—mathematical inflexion points identifiable using the roots of the second derivative—represent the times at which the \( f_0 \) contour appears to be under least pressure to change trajectory. Thus, inflexion points can be taken to represent moments where the contour is least affected by physiological constraints and is most ‘on course’ towards the ideal target. Consequently, the linear slope (or tangent) at the inflexion point can be viewed as an ideal slope towards an ideal target.

This is exemplified in Figure 2, which shows \( f_0 \) and \( f_0''(t) \) contours of a model curve. The blue lines indicate the tangents at inflexion points between times maximum curvature. The intersection of these two lines can be viewed as the unrealised ideal target which would be achieved if physiological constraints and segmental pressure did not cause the speaker to make adjustments to the \( f_0 \) trajectory. Using these tangents and their intersections, an idealised linear interpolation of the contour can be estimated.

![Figure 2: Example of \( f_0 \) contour (black line), its second derivative (red line), and linear slopes projected from inflexion points (light blue line).](image)

3. Implementation

This section outlines the implementation of K-Max, in terms of how it manages \( f_0 \) contour correction, estimates turning points, permits user intervention, and generates an idealised and smoothed pitch contour (see Figure 3). It should be noted that all \( f_0 \) processing is carried out using semitones re 100 Hz.

A corrected \( f_0 \) contour is generated which is then smoothed and used in the rest of the analysis, in an approach similar to
that used by the IPO in generating stylized contours [20]. The corrected contour is produced via the procedure @fixPitch. This exploits the ‘To Manipulation’ function of Praat and provides the option for the user to correct the original contour by removing segmental effects not associated with intonation—such as those caused during voiced fricatives or at the onset of voicing after voiceless stops—and to correct pitch tracking errors such as pitch halving and pitch doubling. To facilitate the estimation of \( f_0(t) \), the corrected pitch contour is interpolated and smoothed using Praat’s in-built functions.

\[ \text{Figure 3: Flowchart of the analysis, annotation, and processing for a single utterance.} \]

After this, the procedure @k is called, which estimates times of \( K_{\text{max}} \) and \( K_{\text{max}} \) using \( f_0(t) \). This includes near-roots, which are points where \( f_0(t) \) approaches but does not reach zero. Even after smoothing, the script will likely identify more \( K_{\text{max}} \) time points than there are salient tonal events. Conversely, as a result of over-smoothing, \( @k \) may occasionally identify too few turning points. User intervention during the procedure @mainUI helps deal with such problems.

To resolve the problem of multiple \( K_{\text{max}} \) time points, the user is prompted to annotate a tonal tier using only turning points which appear salient (see Figure 4). This includes the obligatory annotation of boundaries, but otherwise, the user is free to use any annotation convention they see fit here. If there are too few \( K_{\text{max}} \) time points, the user can adjust the Praat smoothing parameter in the user interface and then re-smooth the \( f_0 \) contour for processing using the new smoothing parameter.

During the user intervention stage, the picture window always displays the pitch contour and a single target tier. \( f_0 \) is indicated on the y-axis while shape size and colour intensity are used to indicate Cepstral Peak Prominence (see Figure 5). This helps distinguish more periodic components of the contour from less periodic ones (such as during voiced frication) and thus identify those parts of the contour which may be more relevant to intonation [21], [22]. The corrected contour can also be displayed in the picture window, so if there appear to be errors in the corrected contour, the user can re-run @fixPitch.

\[ \text{Figure 4: Text grid showing estimated turning points ('MaxK' tier) and those marked salient ('tones' tiers).} \]

Once the salient turning points have been identified and annotated, the idealised contour is estimated by the procedure @idealise, which uses the principles set out in section 2.3 above to estimate ideal slopes and targets. This procedure also deals with several problems which may arise during estimation. First, there will be instances with more than one inflexion point between two turning points. Therefore, @idealise identifies the first and last roots of \( f_0 \) between two extrema. It then performs a linear regression between these points on the \( f_0 \) contour to calculate the ideal \( f_0 \) slope between the two turning points. Conversely, there may be no inflexion points between turning points, such as occurs in plateaus (see Figure 1), in which case near-roots are used. If there is neither a root nor near-root, which can occur at boundaries, slope is calculated using the two edge-most frames. These slopes are used to identify ideal targets. If the procedure can still not adequately generate ideal targets, the user is warned to make adjustments.

Using ideal targets and slopes, an idealised \( f_0 \) contour is generated. Since this contour does not account for physiological constraints, it tends to sound like a much more exaggerated version of the original. At present, to simulate the effects of physiological constraints, a triangular moving point average (MPA) smoothing function is passed across the idealised contour. The width of the MPA size (in frames) can be changed manually in the main UI window in order to ensure that a reasonable approximation of the original contour is achieved.

Once the smoothed idealised \( f_0 \) contour has been generated, the utterance is resynthesized and the picture window is updated. Using visual and auditory judgments, the user can decide if the location of the ideal targets and the resynthesis are acceptable. Finally, the updated text grid is stored along with tables containing \( f_0 \) and time data to facilitate statistical analysis. This includes text grid annotation, time, and \( f_0 \) of turning points as well as the data regarding the ideal contour, i.e., ideal time points, \( f_0 \) values and slopes.

3.1. Testing and Effectiveness

A test set of 83 utterances was analysed and processed using Kmax by the author. There were 7-8 utterances from 11 speakers of northern Irish English (5M, 6F). The pitch accents of the utterances had previously been analysed by the author and another trained phonetician using IvIE annotation conventions [23], but the original analyses were not consulted during this process. Since the quality of the resynthesis is dependent on the
number of turning points specified by the user, the author attempted to mark only turning points which appeared salient as edge tones, pitch accents, and the edges of plateaux / valleys.

As K-Max involves a semi-iterative process which allows the user to test and retest output visually and auditorily and to adjust smoothing parameters on the fly, the results tended to be very satisfactory. That is, it was always possible to identify salient turning points from the ‘maxK’ text grid tier (see Figure 4) which agreed with the author’s intuitions. Furthermore, the final resynthesized output tended to sound practically indistinguishable from the original. In only two cases was it necessary to include a turning point which was not readily identifiable as belonging to the prescribed categories. In each case, these were in nuclear accents occurring in feet with four syllables. This may be a fault of the underlying approach or may reflect the weakness of using linear interpolation between ideal targets.

So far only a small impressionistic test of the idealised and smoothed resynthesis has been conducted. 11 utterances, one each from each speaker, were selected randomly. Each original and resynthesized version was played to four colleagues at the Speech and Phonetics Laboratory. Two listeners judged all acceptable, while two judged one each to have boundaries which were slightly different from the original (for example, see Figure 6). However, in these cases, they felt that this did not affect the overall interpretation of the contour.

4. Applications

As intended, the data provided by K-Max can be used in standard AM analysis of temporal alignment and scaling of tonal targets. However, applications of K-Max stem from the fact that, to create a reasonable resynthesized pitch contour, it requires over-specification of turning points when compared with typical AM analysis. This consequence was intended by design and its benefit will be exemplified with two short illustrations.

Pierrehumbert’s original analysis of downstep [10] argued that it was the obligatory result of a sequence of H and L tonal targets; however, Ladd critiqued this [24], observing that in her data, L targets were sometimes added ad hoc in order to justify such an analysis. Ladd felt the problem boiled down to the difficulty in “reconciling goals of phonetic specification and linguistic generalisation” [24, p. 725]. If one considers the contour in Figure 5, there is a clear sequence of down-stepped H* pitch, which can be represented in AM phonology as H* !H* !H*L L% (following IViE annotation conventions, [25]). Yet in order to be able to resynthesize the contour adequately, it is necessary to annotate turning points at the edges of the down-stepped plateaux. The left edge of each down-stepped plateau has been annotated as L_ to show the turning point is associated with the upcurrent H* and is not the tail of the previous pitch accent. In essence, this surface L*H* sequence can be viewed as the phonetic implementation of an underlying H* pitch accent without compromising the phonology. In fact, it provides a means by which one might be able to identify more precisely the mechanisms through which the underlying phonology is implemented, or, in other words, reconcile phonetic specification with linguistic generalisation.

As a second example, K-max can be used to quantify and analyse plateau boundaries. In Figure 5 and Figure 6, the right edges of the plateaux have been annotated with _0. Again, the underlying phonology is still evident (L*H L*H %. In the latter case), but the inclusion of _0 turning points is needed for resynthesis. In each utterance, _0 occurs at or near the right edge of the stress-containing word. If nothing else, this hints at a need for a more detailed analysis of the alignment of plateau edges and their potential role in signalling lexical boundaries.

![Figure 5: Example of output from the picture window showing pitch contours and tonal targets.](image)

![Figure 6: Original and smoothed idealized contour of an utterance from a northern Irish English speaker.](image)

In short, the inclusion of a resynthesis component forces the user to identify a minimal number of turning points which are required for contour realisation. This, in turn, encourages the user to consider the role of such turning points and their relationship to the underlying phonological forms.

5. Conclusions

This paper has outlined the rationale behind an AM-based intonation analysis tool, K-Max. It identifies turning points using $f_0(t)$ and resynthesizes the contour using ideal tonal targets and slopes along with a simulation of physiological constraints on $f_0$ variation. It is argued that the inclusion of a synthesis component, while not typical of an AM approach, helps draw attention to and facilitate the analysis of intonationally significant components of the $f_0$ contour which are under-analysed in the AM approach.

Refinements will be made to the physiological constraints function and further objective and subjective tests of the resynthesis will also be conducted. Moreover, K-Max will be used to analyse larger corpora, which will provide more quantitative data regarding its effectiveness, and hopefully demonstrate its usefulness.
6. References


