Smooth Contour Estimation in Data-Driven Pitch Modelling

Kim E.A. Silverman*  Jerome R. Bellegarde*  Kevin A. Lenzo†

* Spoken Language Group
Apple Computer, Inc.
Cupertino, California 95014
{kimsilv, jero me}@apple.com
† School of Computer Science
Carnegie-Mellon University
Pittsburgh, Pennsylvania 15213
lenzo@cs.cmu.edu

Abstract

Apple’s next-generation text-to-speech (TTS) system in Mac OS X uses a superpositional pitch model, comprising a relatively smooth underlying F0 contour and a separate contribution from the influence of the phonetic segments. This paper focuses on the data-driven modelling of the underlying contour, based on electroglossographic signals obtained from a corpus of reiterant speech. F0 extraction from such signals leads to more accurate characteristic shapes, as objectively illustrated by a typically low mean absolute frequency deviation (between 2 and 3 Hz) between original and synthetic F0 contours. This in turn supports a better (both more complete and more realistic) model of F0 behavior. Experimental results illustrate the improved prosodic representation resulting from this F0 model.

1. Introduction

This work concerns the automatic generation of fundamental frequency (henceforth F0), the primary acoustic correlate of pitch in speech. F0 modelling involves the capture of both segmental and suprasegmental information. Overall F0 variation is determined by the prosodic structure, and thereby conveys information about the semantic role and dialogue function of sentences—information that cannot be derived from the words themselves. At the same time, the phonemes that make up the words strongly influence the more local behavior of F0 variation, and thereby perturb the fundamental frequency values away from the underlying contour [1].

Accordingly, we model F0 variation as a superposition of relatively local segmental perturbations and the underlying suprasegmental structure of intonation (cf. [2], [3]). This approach is different from other suprasegmental models (e.g., [4], [5]) in that the superposition here is strictly linked to the segmental level. Thus, for example, we do not further decompose the underlying contour into accentual tones riding on top of phrasal tones.

The suprasegmental structure of F0 comprises locally-defined and sparsely-distributed pitch accents, phrasal tones, and the overall pitch range in which these occur. They jointly produce a relatively smooth contour, which can be thought of as commands sent to the larynx. This paper focuses on the characterization of this smooth contour. In contrast with decision tree methods (e.g., [6]), we rely on the ToBI transcription system [7], [8]. This leverages agreement across the major traditions of intonational analysis, and allows us to directly relate the transcription to typical text processing performed in the front-ends of synthesizers and dialogue systems.

With a ToBI representation of the intended intonation [8], speech F0 contours are specified by the choice and locations of the pitch accents, phrase accents and boundary tones. Our model parameters are statistically estimated from electroglossographic (EGG) signals in a corpus of reiterant speech, described in [9]. All the parameters are estimated from a single speaker, speaking in a single and consistent style, within a single recording setup.

The paper is organized as follows. The next section briefly reviews the mechanics of generating F0 contours from a ToBI transcription. Section 3 discusses the smoothness of the contours so obtained. In Section 4, we describe an improved approach to contour estimation. Finally, Section 5 presents some experimental results to illustrate the associated benefits.

2. ToBI-Derived Contours

A common approach to generating F0 contours from a ToBI-like transcription was first presented in [10], and is illustrated in Fig. 1. Each component of the accent inventory is modelled as an abstract shape. For example, final boundary tones consist merely of a plateau, while H accents consist of a plateau preceded by a vertical “leg,” resembling the uppercase Greek letter “ơ.” On a time/frequency plot, these shapes are aligned with their associated syllables (Fig. 1a), and then any gaps between them are linearly interpolated (Fig. 1b). Then the contour is causally smoothed by convolution with a rectangular window, whose duration is equivalent to the length of the plateaux (Fig. 1c). This equivalence is necessary in order to ensure that the smoothed contour reaches the target values of the accents. The resulting filtered concatenation of straight-line segments is intended to integrate locally-invariant accent shapes with global, contextually-determined slopes and heights of interpolations between the accents.

Silverman [3] modified this approach in a number of ways. The relative heights of fundamental frequency

*This work was performed while Kevin Lenzo was with the Spoken Language Group at Apple Computer.

†Many thanks to Victoria Anderson: without her linguistic skill and attention to detail, this study would have had no data.
events were scaled exponentially, rather than linearly. This is psychoacoustically more defensible, because F0 differences are perceived in logarithmic increments. Linguistically, such scaling allowed H and L tones to be treated symmetrically: increasing local emphasis on a pitch accent could be implemented the same way independently of the tone type. And non-linear scaling did away with the need for phrase-final raising in some question contours.

Another modification concerned the window shape. Rectangular windows often produce unnaturally sharp peaks and valleys, and sudden changes in the direction of F0 contours, as illustrated in Fig. 1.c. Production of such discontinuities in F0 or its derivatives from a human larynx would require brief moments of extremely high muscular force, and sudden extreme changes in muscular force. Studies of human motor control often characterize muscular movement in terms of peaks of acceleration and rate of change of acceleration (known as “jerk”.) Nelson [11] found that human speakers economize on effort by minimizing jerk, and that the minimum-jerk velocity profile for muscular movement is almost indistinguishable from simple harmonic motion. Accordingly, the rectangular window was replaced with a Hamming window. This minimizes all higher derivatives, approaching the pitch accents in a way that reduces jerk and hence speaker effort. Yet the targets are still guaranteed to be reached, albeit in a smoother way (Fig. 1.d).

3. Smoothness Issues

Still, there are a number of problems associated with the use of a fixed-duration window to generate the relatively invariant local shape of pitch accents and the overall global interpolation between pitch events. At best, this can only be partially successful, because the window cannot satisfactorily handle both the local shapes and the necessary smoothness of the contours. These issues are illustrated in Figs. 2 and 3.

Fig. 2 shows EGG-derived F0 contours for two common phrase-final tunes: (i) a nuclear fall H* L L%, and (ii) a nuclear full-rise H* L H%. Since these tunes are associated with the most perceptually and semantically salient words, it is important to synthesize them well. This is difficult, however, when they are spread over different lengths of segmental material and different number of words. In Fig. 2 it is clear that these two different tunes have much in common. Both have an F0 peak on the accented word, followed by a steep fall, no matter how far to the end of the utterance. (In the figure, all of the contours are aligned by this peak.) The primary difference between the tunes is the presence or absence of a final rise corresponding to H%. This is appearable to being modelled by a sequence of a low and high targets aligned with the end of the utterance, independently of the nuclear H* to L transition. While a cursory examination of these data implies a locally invariant characteristic shape for pitch accents, more detailed analysis shows that the F0 movements have a variable steepness which cannot be rendered well by a fixed-duration window. For example, rises tend to be less steep than falls, and also less steep in prenuclear than nuclear position. And falls themselves tend to be less steep when the following L% is located some distance further downstream in the utterance. The effect seems to be that speakers guide some effort to ensure that target F0 levels are reached, and so will move their F0 more steeply when there is less segmental material to carry the movements, but will relax their effort a little when there is more segmental material available.

Fig. 3 shows EGG-derived F0 contours for the tone sequence H+L* H* L L%, aligned by the H* to L transition. The utterances differ in the number of unaccented syllables separating the H+L* from the following H*. Clearly, the final nuclear sequence can be modelled independently of the preceding material. (Similar plots aligned by the H+L* accent show that its local shape, too, is contextually invariant.) Also strong in Fig. 3 is evidence that F0 follows an almost linear interpolation between the two accent gestures, with all the contours converging at the start of the small local rising gesture that is the beginning of the final H* (cf. phrase 378). Again, this argues in favor of a global interpolation between pitch events with relatively invariant local shapes. A fixed-duration window tuned for fast utterance-final
falls, however, will tend to either approach prenuclear accent peaks too steeply, or fall from nuclear accents too slowly.

4. Refined Estimation

Because of such difficulties, some authors have opted to replace characteristic shapes with approximation functions (cf., e.g., [12], [13]). For example, if F0 movements are assumed to conform to a fourth degree polynomial variation, the parameters of the associated function can be easily trained using, e.g., neural techniques [12]. Note, however, that in this model, global movements between pitch accents are still interpolated, so window optimization is still an issue, albeit implicit.

In the present approach, we keep the basic framework of parameterizing each EGG-derived F0 contour by a number of characteristic shapes related to ToBI symbols, and using linear interpolation to fill gaps between characteristic shapes. However, the control points associated with the pitch events are placed more carefully, for example by relaxing the constraint that the “leg” preceding a tone plateau be vertical. With the abstract accent shapes thus taking more responsibility for the local accent shape, the smoothing window is required to do less work and so the smoothing window can contract. Since each characteristic shape can be described by a relatively small number of local targets (typically 2 or 3), we can use gradient descent to jointly optimize the abstract shapes of the underlying accents and the window duration. Thus, a single set of parameters is trained across the entire corpus.

The contours were hand-marked with a ToBI transcription, indicating the peaks corresponding to H tones and the valleys corresponding to L tones. In bidental accents, such as L+H, the alignment point was the starred tone. The gradient descent optimization was then able to adjust these alignments to improve the overall fit. This makes sense because the precise locations of local maxima and minima in the extracted F0 contours are subject to some quantization and variability. The optimized loc-

ations take into account the surrounding contour as it approaches and leaves the turning points.

Each monotonical accent was represented as a plateau (i.e., two target points with the same F0 value) preceded by one extra target point. The rightmost of these points was aligned with the corresponding F0 target, and the duration of the plateau corresponded to the duration of the smoothing window in order to guarantee that the target F0 would be reached. In contrast with prior work, the duration of this plateau was jointly optimized with the rest of the model parameters. Each time the optimization algorithm iteratively adjusted the duration of the smoothing window, the leftmost point of the plateau was adjusted accordingly. Also unlike prior work, the first target point of the accent was not constrained to be placed in the frame immediately before the beginning of the plateau, but was free to vary in time and frequency, with two constraints: (i) it could not be located higher in frequency than the associated following H or lower than the associated L, and (ii) it could not occur earlier than any immediately-preceding pitch event.

5. Experimental Results

F0 modelling quality was objectively assessed on held-out portions of the underlying corpus of reiterant speech (see [9] for a complete description of this corpus). The evaluation criterion was taken to be the mean absolute frequency deviation across all utterances considered. Note that this is different from usual criteria like the root mean square (RMS) error and the correlation ($R^2$) between original and synthetic contours. While such measures have been used in a number of information evaluations (e.g., [14]), and were shown by Hermes [15] to be well motivated similarity metrics, we believe the mean absolute frequency deviation may be even better suited for determining how well the new F0 models capture the qualities of the underlying data. This is because RMS and $R^2$ are both based on the $L_2$ norm, which tends to reduce the effect of rare outliers. In F0 modelling, rare outliers can be perceptually egregious, and we therefore want to avoid downweighting their influence.

Fig. 4 shows the result of gradient descent optimization on one of the F0 contours shown in Fig. 3. The solid line interpolates between the targets, and then undergoes causal smoothing with a Hamming window. Fitted fundamental frequency values are generated in 5 ms frames. In this example, the mean absolute deviation between fitted and original values is about 1.9 Hz. Over the entire corpus, this figure typically varies between 2 and 3 Hz.

Fig. 4 illustrates several characteristics of the approach. One is that most of the model error arises from the segmental perturbations remaining in the EGG signal (see, for example, frames 50 to 50). This confirms that the residual from the fitted smooth contours will provide good data for systematically modelling the perturbations. Another is that the optimal smoothing window (60 ms long) turns out to be considerably shorter than in prior work (e.g., 180 ms in [3]). Consequently, the local accent targets are more critical to the generation of the smooth curves. For example, an extra target (at frame 80) was needed between the high plateau at the start of H+L (at frame 67) and the low plateau corresponding to L (at frame 118). This extra target became
part of the abstract characteristic shapes for all H+L* accents, and illustrates yet another way in which the local shape around the starred tone cannot be modelled by the simpler approach of prior work.

The data argued that the abstract shape values for accents depend on the context in a number of ways. For example, the height of the first target of an H* depends on the identity of the preceding tone. Specifically, when preceded by a H+L*, as in Fig. 4, it needs to be about 15% lower than when preceded by L, H*, or L+H*. In another case, the location of the L phrase accent after an immediately-preceding nuclear H* depended on the linear distance to the end of the associated intermediate or intonational phrase. In general, model parameters for most accents depend on the phonological identity of the immediately-preceding event on the tonal tier, regardless of its distance, and the phonetic detail (timing) of the following event on the tonal tier, regardless of its phonological identity. We note that this is the opposite of phonetic anticipatory co-articulation, where the articulation of a phoneme depends largely on the identity of the following phonemes. Thus, for example, the /k/ in "carp" has a different place of articulation than the /k/ in "keep."

Finally, note that the first target occurs 40 ms before the start of the utterance. The data showed that when modelled this way, the F0 contour across unaccented utterance-initial syllables could be modelled with one initial value across the entire corpus. This holds regardless of whether it subsequently rises to a H tone or drops to a L tone on the first accent, and regardless of the distance to that accent.

6. Conclusion
We can make progress in modelling F0 by generating an underlying smooth contour which is then perturbed by the phonetic segments. An EGG signal obtained from a corpus of relevant speech supplies a good estimate of this underlying contour. Such an estimate in turn supports accurate data-driven modelling, as objectively illustrated by a typically low mean absolute frequency deviation (between 2 and 3 Hz) between original and synthetic fundamental frequency variations. This translates into a better synthetic reconstruction of the underlying smooth contour.

In particular, we moved the responsibility for generating the relatively-invariant local shapes from the smoothing window to a more careful placement of the targets that make up those shapes. This placement is conditioned on the identity of and distance to neighbouring phonological pitch events. The resulting model of F0 behavior is therefore more complete and more realistic. We are currently gathering large-scale subjective evidence to confirm that the improved prosodic representation resulting from such F0 model leads to more natural-sounding synthetic speech.

7. References