MODELLING AND SYNTHESISING F0 CONTOURS WITH THE DISCRETE COSINE TRANSFORM

Jonathan Teutenberg[1]*, Catherine Watson*[2]*, Patricia Riddle* [1]

[1]Department of Computer Science
[2]Department of Electrical and Computer Engineering
University of Auckland

ABSTRACT

The Discrete Cosine Transform is proposed as a basis for representing fundamental frequency (F0) contours of speech. The advantages over existing representations include deterministic algorithms for both analysis and synthesis and a simple distance measure in the parameter space. A two-tier model using the DCT is shown to be able to model F0 contours to around 10Hz RMS error. A proof-of-concept system for synthesising DCT parameters is evaluated, showing that the benefits do not come at the expense of speech synthesis applications.

Index Terms— Discrete Cosine Transform, speech processing, speech synthesis

1. INTRODUCTION

As part of a long term study into voice modification, we intend to alter the prosody of existing speech by modifying the F0 contour. This modification will achieved through the adaptation of an F0 model.

A number of models for F0 contours of speech have been proposed and are in use ranging from high-level phonological models [1] to surface representations of the contour (e.g. [2, 3]). Regardless of the form a model takes, there are three important processes that need to exist to enable its use in voice modification. The first process we shall refer to as *analysis*, which obtains model parameters from an F0 contour (sometimes referred to as coding). The second is the inverse process of *F0 synthesis* that produces an F0 contour from model parameters. The third process is *parameter synthesis*, producing model parameters from lexical features of an utterance - a process typically used by speech synthesis systems.

An ideal model of F0 contours would have deterministic implementations of the above three processes. Other desirable properties include language independence so that a single method can be developed for all languages in parallel, a scalable trade-off between precision and model complexity for low-bandwidth applications, and a distance measure for the parameter space to allow clustering and codebook generation as in [4].

Perhaps the most widely used model for F0 contours is ToBI, a symbolic representation at the phonological level [1]. Mature algorithms for both F0 synthesis and parameter synthesis exist, based on statistical methods learnt from data [5]. A downside of the use of statistical methods for these processes is that the parameters obtained from a given contour are dependent on the data used in training, so there is no guarantee of consistency between the parameters generated from two implementations of a process based on differing data. Furthermore, a statistical method's output on data that varies greatly from its training data (such as a new speaking style or a new language) is undefined. A further limitation of ToBI is that as yet no automated analysis algorithms exist and corpora must be annotated by hand by an expert.

Models that lie closer to a surface representation of the F0 contour possess fewer limitations than ToBI, such as Fujisaki's model [2], Tilt [3] and various others [4, 6, 7]. These models were defined with a focus on F0 synthesis and provide succinct, deterministic algorithms for this process. The analysis processes all rely on some form of event detection, involving statistical methods [3, 6, 8]. Parameter synthesis is also performed using statistical methods [4, 5, 6, 7, 9]. As all of these methods rely on the notion of events appearing at unrestricted points in time, individual parameters are not comparable and no trivial distance measures exist for their parameter spaces. In addition, these methods have a fixed number of parameters which cannot be reduced for concise storage or transmission.

In this paper we propose an F0 model that has a deterministic algorithm for both analysis and F0 synthesis, provides a simple distance metric and allows scaling for low-bandwidth transmission. Furthermore, we show that this model can be used in parameter synthesis methods as in [5, 6, 7, 9] by creating a simple synthesis model using decision trees.

2. DISCRETE COSINE MODEL

For voiced speech, the F0 contour varies slowly and continuously over time. It is therefore well modelled by a half cosine and its multiples, that is, the Discrete Cosine Transform.

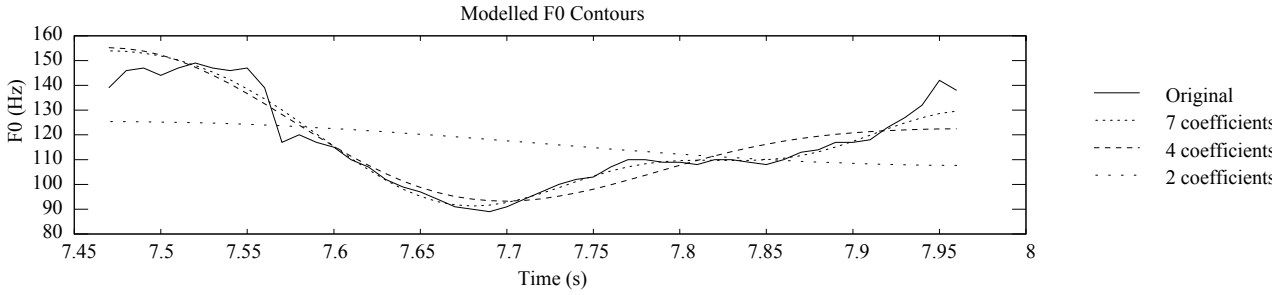

Fig. 1. Modelling the voiced segment *"the one who"* with 2, 4 and 7 DCT coefficients. The mean (c_0) coefficient is included in the number of coefficients.

The Discrete Cosine Transform (DCT) approximates a finite signal of length M with a sum of weighted cosine functions with zero phase. The first DCT coefficient is the mean of the signal, the second is the weighting for a cosine function with wavelength equal to $\frac{M}{2}$, and each successive coefficient corresponding to a function with wavelength $\frac{M}{2}$ higher than its predecessor.

$$s(x) = \tfrac{1}{2}c_0 + \textstyle\sum_{n=1}^{N-1} c_n cos\left(\frac{\pi}{M}n(x+\frac{1}{2})\right) \quad (1)$$

Equation 1 shows how a signal, s of length M can be decomposed into N DCT coefficients $\{c_0, c_1, \cdots c_{N-1}\}$. This describes the inverse-DCT, used for F0 synthesis. The DCT, which is the equivalent process for analysis is:

$$c_n = \textstyle\sum_{x=0}^{M-1} s(x) cos\left(\frac{\pi}{M}n(x+\frac{1}{2})\right) \quad (2)$$

A model for F0 contours based on the DCT coefficients can be seen as similar to the additive models used in [6, 7], where the contours are modelled by a sum of smooth functions. However the DCT coefficients differ in that rather than being a sum of three or four functions offset in time, they instead are sum of up to M periodic functions of length M.

Figure 1 shows the approximation of a segment of voiced speech using the DCT with $N = 2$, $N = 4$ and $N = 7$. As the number of coefficients used increases, the contour represented by the DCT coefficients rapidly approaches the original F0 contour.

2.1. Two-tier model

In this paper we propose a two level model, similar to those of [2, 6], where one set of parameters is used to represent the overall phrase level contour and further sets of coefficients are used to represent the detail of each voiced segment.

The phrase-level coefficients are used to represent a contour that passes through the mean of each voiced segment, giving the overall cadence of the phrase. The first coefficient (the mean) of each voiced segment can therefore be omitted, with the remaining DCT coefficients representing micro-variations in the contour such as those due to syllable stress.

The analysis process for obtaining DCT coefficients from an F0 contour first takes the DCT of each voiced segment using Equation 2. As no F0 contour can be calculated for unvoiced segments, these are then filled by linear interpolation. Next, the DCT coefficients are computed for the phrase.

F0 synthesis is the inverse of the analysis process. First the mean of each voice segment is determined by taking the inverse-DCT of Equation 1 for the phrase level DCT coefficients, then the same is performed for each voiced segment. Together these form an approximation of the original F0 contour.

The process of parameters synthesis for this model was achieved using a statistical approach. We present this in the following section.

2.2. Distance metric

The DCT model provides a simple distance metric for comparing the similarity of two phrase level components, or two voiced segments. As a DCT coefficient is a weighting of the same sinusoidal component in any two contours, these can be compared directly. If two contours are represented with a different number of coefficients, a comparison can still be made as the missing higher-order coefficients are effectively set to zero.

The Euclidean distance between the two sets of coefficients in the parameter space can be used as a measure of similarity between the shapes of two contours. For most practical applications the relative durations of the contours also needs to be incorporated into the distance metric, as the same contour shape over differing durations may be perceived differently.

2.3. Modelling error

To assess the accuracy at which the DCT coefficients model natural F0 contours, a number of phrases were analysed using the two-tiered model described above, then resynthesised and compared to the original. The speech data used was the Keele database [10] of 10 speakers (five male, five female) reading

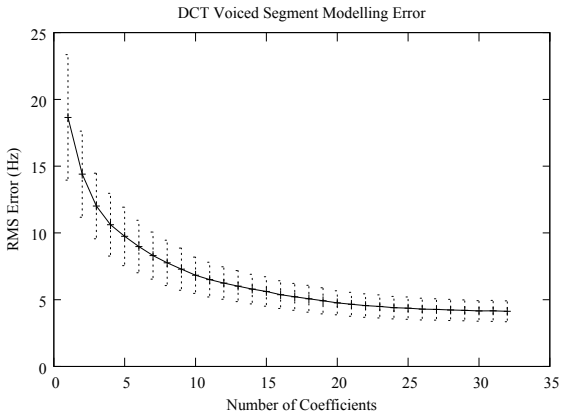

Fig. 2. Modelling error of the DCT over voiced segments using between 1 and 32 coefficients. Error bars indicate one standard deviation above and below the mean error.

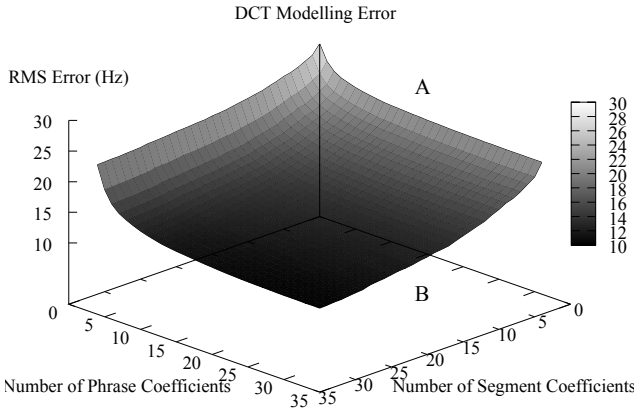

Fig. 3. Modelling error of the two-tiered model over phrases, using between 1 and 32 coefficients for the phrase and for voiced segments. Dark colours indicate low error.

a short story from a set text, which was on average about 30 seconds. The pitch contours provided by this database were recorded by a laryngograph and then hand corrected, so errors due to pitch tracking do not influence our results.

Phrase boundaries were marked by hand, and any continuous series of pitch marks was taken to be a single voiced segment. All combinations from 1 to 32 coefficients for each of the phrase level and voiced segment level DCTs were computed, the F0 contour resynthesised and an RMS error between the orignal and the resyntheised F0 contours were calculated across all 10 speakers.

The error rate on the voiced segments is shown in Figure 2. The RMS modelling error reduces to under 10Hz (6.7%) using 5 coefficients. Additional coefficients reduce the error slighty, to around 5Hz (3.4%) when more than 20 coefficients are used. This is consistent with our earlier work modelling formant trajectories with the DCT [11], where the slowly varying nature of the contour means that few coefficients are required for an accurate model.

Figure 3 shows the modelling error for the phrase level coefficients when combined with the voiced segments in a two-tier model. The most interesting features of the surface shown are the four sides of the quadrilateral. The side labelled A shows that when modelling a phrase without the detail of the voiced segment level, using more than 10 coefficients provides no reduction in modelling error.

The side labelled B in Figure 3 shows that the 5Hz RMS error from modelling the voiced segments (seen in Figure 2) is increased to around 10Hz (6.7%) when extended to include a phrase level component. From the figure it can also be seen that to get within a few Hertz of the minimal error values for the F0 contours in the Keele database, around 6 phrase level coefficients and 10 voiced segment level coefficients are sufficient.

3. PARAMETER SYNTHESIS

To determine whether a DCT representation of F0 contours is still amenable to statistical methods of parameter synthesis, a small experiment was conducted.

The speech of the five male speakers of the Keele database were manually marked up with phrase boundaries and accent locations. Four of these speakers were then used to train decision trees (as used with other F0 models [5, 7]) to predict DCT coefficients at both the phrase and voiced segment level of the two tiered model described in the previous section. At the phrase level 6 coefficients were used, and at the voiced segment level 10 were used. A total of 14 decision trees were learnt, each predicting one of the DCT coefficients. The trees were learnt using the regression tree algorithm of WEKA [12] with default parameter settings.

The success of the parameter synthesis was determined using the remaining speaker. The decision trees were used to predict the DCT coefficients, which were then used to synthesise an F0 contour for the speaker. This contour was then compared against the speaker's original F0 contour.

3.1. Features

The features used to train the decision trees at both voiced segment and phrase level are similar to those of [4, 9, 7]. At the voiced level these were: segment duration, accent location and segment location. The accent location was -1 if no accent occurred within the segment, and otherwise was a number from 0 to 1 given the location of the accent within the segment. The segment location takes one of three values: phrase initial, phrase final or middle.

At the phrase level the following features were used: phrase duration, accent location and phrase location. The accent location, as for the segment level, was a value from 0 to 1 and exactly one accent was specified per phrase. The phrase lo-

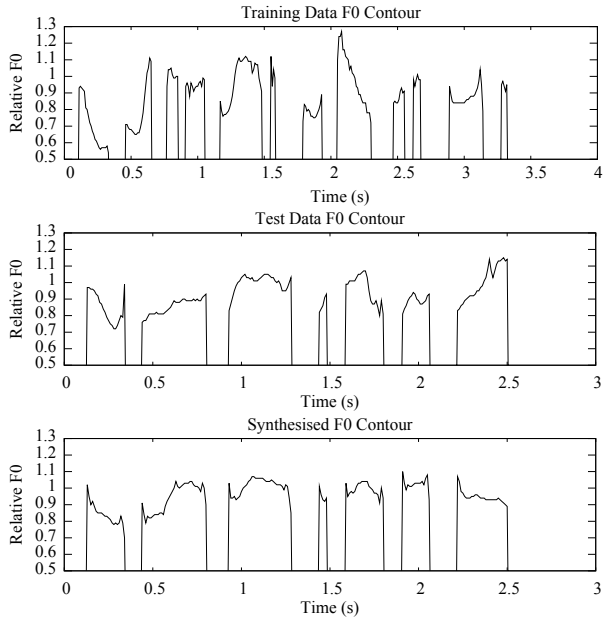

Fig. 4. F0 contours of *"The North Wind and the Sun were disputing which was the stronger"* in one of the speakers used as training data (top), in the test data (middle) and as synthesised from the output of the decision trees (bottom). F0 values are given as fractions of the speaker's mean F0.

cation was a number from 0 to 1 giving the phrase's position within the sentence. The nth phrase in a sentence with N phrases was given a phrase location of $\frac{n}{N}$.

3.2. Results

The accuracy of the parameter synthesis model was pleasingly high. Despite a high variation between training and test data (samples of which are given in Figure 4) the mean error on the test data was only 14.4Hz (9.5%), with a standard deviation of 17.3Hz (11.5%). This compares favourably with around 30Hz error obtained by other approaches [7, 6, 9].

In particular the phrase level coefficients were well predicted, with most of the error due to a few gross differences in voiced segment prediction. One example of these gross errors is in the final voiced segment in Figure 4, where the original speaker used a rising contour but a falling contour was predicted. Of course, these differences do not necessarily indicate that either is an unnatural contour. Applying the predicted contour to the original speech and resynthesising using overlap and add resulted in a plausable utterance.

4. CONCLUSION

A new representation of F0 contours based on the DCT has been proposed. Unlike other representations, simple, deterministic algorithms exist for both analysis and F0 synthesis. Contours can be easily clustered using the Euclidean distance between sets of coefficients. It has been shown that with a two-tiered approach a contour can be modelled with a high level of accuracy using relatively few parameters.

The production of a proof-of-concept parameter synthesis model shows that the benefits of a DCT representation do not come at the expense of its applicability to speech synthesis.

This makes a compelling case for the use of the DCT in the area of F0 contour modelling.

5. REFERENCES

[1] K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, and J. Pierrehumbertand J. Hirschberg, "TOBI: a standard for labeling english prosody," in *ICSLP-1992*, 1992, pp. 867–870.

[2] H. Fujisaki and S. Ohno, "The use of a generative model of F0 contours for multilingual speech synthesis," in *ICSP-1998*, 1998, pp. 714–717.

[3] P. Taylor, "The tilt intonation model," in *ICSLP-1998*, 1998, pp. 1383–1386.

[4] T. Kagoshima, , M. Morita, S. Seto, M. Akamine, and Y. Shiga, "An F0 contour control model using an F0 contour codebook," *Systems and Computers in Japan*, vol. 38, no. 1, pp. 976–986, 2007.

[5] A. Black and A. Hunt, "Generating F0 contours from ToBI labels using linear regression," in *ICSLP-1996*, 1996, pp. 1385–1388.

[6] S. Sakai, "Additive modeling of english F0 contour for speech synthesis," in *ICASSP-2005*, 2005, pp. 277– 280.

[7] X. Sun, "F0 generation for speech synthesis using a multi-tier approach," in *ICSLP-2002*, 2002, pp. 2077–2080.

[8] H. Ogawa and Y. Sagisaka, "Automatic extraction of F0 control parameters using utterance information," in *SP-2004*, 2004, pp. 447–450.

[9] K. Dusterhoff, A. Black, and P. Taylor, "Using decision trees within the tilt intonation model to predict f0 contours," in *Eurospeech '99*, 1999, pp. 1627–1630.

[10] E. Plante, G. Meyer, and W. Ainsworth, "A pitch extraction reference database," in *Eurospeech '95*, 1995, pp. 837–840.

[11] C. Watson and J. Harrington, "Acoustic evidence for dynamic formant trajectories in Australian English vowels," in *JASA*, 1999, vol. 106, pp. 458–468.

[12] Ian H. Witten and Eibe Frank, *Data Mining: Practical machine learning tools and techniques, 2nd Edition*, Morgan Kaufmann, San Francisco, 2005.