OBJECTIVE METHODS FOR EVALUATING SYNTHETIC INTONATION

Robert A.J. Clark and Kurt E. Dusterhoff

Centre for Speech Technology Research, University of Edinburgh,
80 South Bridge, Edinburgh EH1 1HN
http://www.cstr.ed.ac.uk
email: {*robert, kurt*}*@cstr.ed.ac.uk*

ABSTRACT

This paper describes the development and evaluation of objective methods for testing synthetic intonation. While subjective methods are available for assessing the quality of synthetic intonation, such tests consume time and resources, and are not useful for day-to-day model development. Therefore, objective measures of F0 modelling are necessary. Currently, objective evaluation of synthetic intonation involves the use of Root Mean Squared Error and Correlation. However, it is unclear how large an improvement in either score must be before it is reflected perceptually. It is also unclear how detailed an analysis these metrics provide. Therefore, two other metrics are to be tested, both of which are similar to a basic RMSE measurement. All of the evaluation results are compared to a perceptual study in order to determine how the objective measures relate to perceived differences in the contours.

1. INTRODUCTION

One difficulty in building models for synthesizing intonation is determining how well those models capture the qualities of the data which they model. Analysis-by-synthesis methods (e.g. [5]) allow comparisons between an original utterance and one in which only the F0 has been changed. When no difference between the two is perceived, the model can be said to successfully reflect the nature of the data. However, comparing pairs of utterances for a large database is prohibitive. One finds that listening to the pairs for 100 utterances can result in all intonation sounding basically the same. Only testing a few utterances makes it difficult to determine how robust the models are. These obstacles to subjectively testing intonation models make finding objective evaluation methods a worthwhile pursuit. This paper presents an investigation into three methods of objectively (and automatically) judging the similarity of two fundamental frequency contours. The three methods are described in detail, and are related to perceptual judgements in order to determine which method best matches listeners' perceptions.

The first method for evaluating F0 differences involves using RMSE and Correlation measurements [2]. The basic RMSE measures the distance between two contours on the time axis. High RMSE scores show a large difference in F0 between the two contours, while low RMSE shows a small difference. Correlation coefficients reflect how well the direction of the synthetic F0 follows the original. When both the original and the synthetic contours are rising , for example, the coefficient is higher than if one is rising and one is falling. These measures have been used in a number of intonation evaluations (e.g. [4], [1]), and were shown by Hermes [3] to be the best currently used similarity metrics.

The two assessment methods which will be tested against this basic measurement attempt to combine time and frequency evaluation into a single metric.

First to be tested is an RMSE measurement of the distance, on lines normal to the the point on the reference F0 curve. These lines correspond to a component in the direction of the acceleration of the curve.

The second metric divides the F0 curves into sections between "anchor points", which can be related to pitch events. The distortion between each section of the test and reference curve between these anchor points is measured by dividing each section into a set number of equally spaced points, with respect to curve length, and finding the RMSE distance between them.

In conjunction with the objective measurement techniques listed above, subjects were asked to rank pairs of utterances on a scale of 0 to 4, where 0 reflects no audible difference in intonation, and 4 reflects no audible similarity. The subjects rated 24 utterance pairs. The subjective ratings were normalized, and a Pearson test was run to discover whether any of the objective metrics correlated with the subjects' perceptions.

2. THE PERCEPTUAL TEST

The subjective evaluation of intonation difference consisted of asking novice subjects (first and second year undergraduate students) to rate utterance pairs on a five-point scale. The pairs, bar a control set, consisted of one stimulus synthesized with the F0 which was extracted from the original utterance (and smoothed) and one synthesized from an F0 generated using statistical models ([1]). Four pairs contained only one or the other for both stimuli. The stimuli were generated using LPC resynthesis of the original waveform and the imposed F0. The utterances were presented to the subjects via a web interface over Sennheiser headphones using standard audio software on Sun Ultra workstations. The subjects participated in the experiment in a quiet, closed computing laboratory, and were unable to hear normal levels of ambient noise.

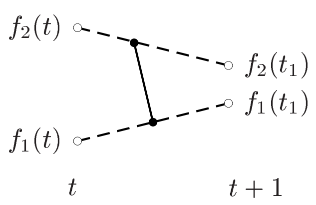

Figure 1: Computing the tangental metric

Nineteen native speakers of British English took part in the test. Four of the subjects produced results which suggested they were unable to adequately perform the task (pairs which were audibly very different were judged on the same end of the scale as pairs which were exactly the same), and their results were removed from further testing. The subjects had a general understanding of intonation when questioned, were provided with written instructions, and were able to practice using the interface and ask questions prior to beginning the test.

The web-style interface consisted of four pages, each with six stimulus pairs and written details about how to use the rating scale and an introduction page with full instructions on their task and seven example pairs. The example pairs also included example scores to illustrate a rough guide to the audible differences which the subjects might hear and as an example of how the web interface worked.

For each stimulus pair, two buttons could be "clicked" by the subjects - one to play each stimulus. Five ranking buttons were line up to the right of the stimulus buttons, with the ranking value heading each column. The subjects could listen to each stimulus as many times as they wanted before chosing a ranking on the scale. Having made the decision, the ranking was selected by "clicking" on the appropriate ranking button.

Subjects who were able to accurately place the control pairs in the 0 or 1 ratings and used at least four of the five possible rating choices were included in the correlation with the objective measures. Those who did not consistently place the control pairs in the 'most similar' range, or who placed all utterances in the 0-3 range were not included, as they did not show a potential to make the sort of fine distinctions that the objective measures are being tested for.

3. ALTERNATE METHODS

One of the drawbacks of the standard RMSE method is that it computes the error between two F0 values at given time points, where intuitively a method which computes errors in pitch-event alignment both with respect to time and F0 would seem more ideal. The methods proposed here are a step in that direction.

An inherent problem with any such metric is that pitch and time are measured on independent axes with different units which cannot necessarily be combined in a straight forward manner. That is, an error of 1 unit on the time axis does not necessarily correspond to an error of 1 unit on the frequency axis. We therefore need to appropriately weight each component when combining them. This study weights 10ms time displacement equal to 1Hz frequency displacement. Heuristic experimental calibration of weights shows that this weighting is the right order of magnitude, but it should only be regarded as a reasonable starting point. Further experimental research would be needed to finely tune this parameter. There are of course also the usual issues of whether frequency should be measured on a linear scale of not, which will not be discussed here.

Two alternate methods to compare F0 contours are presented here. Both methods attempt to utilise the notion that differences between the contours can be brought about by either pitch displacement errors, time displacement errors or a combination of both. The *tangental estimation method* estimates a distance perpendicular to the direction of change of the contour and the *warping method* measures the distance between contour points within a pitch-event.

3.1. Tangental Estimation Method

This method treats one contour as a reference and compares the other to it. The distance between the contours is measured in a direction normal to the reference contour. This is to incorporate a time component based on the assumption that the contours rates of change are similar, as well as the contours themselves being similar.

To calculate these distances, the contours are estimated locally by straight lines. These lines are constructed by taking the data from two consecutive frames and using these time and F0 values as end points for the lines. Then a point is chosen on each line and the distance between the points is taken as the distance between the contours. On the reference line, the midpoint between frames is chosen. On the comparison line, the intersection of the comparison line and the line normal to the reference line passing through the previously chosen point is selected.

If $f_1(t)$ and $f_2(t)$ are the reference line and comparison line respectively, estimated from the contours across frames t and $t+1$ (as shown in Figure 1), then the squared distance between the contours is calculated (by translating the time axis to the origin) as:

$$d^2 = (r_x - c_x)^2 + (r_y - c_y)^2. \tag{1}$$

where:

$$c_x = \frac{1}{2}\frac{[f_1(t+1) - f_1(t)][f_1(t+1) + f_1(t) - 2f_2(t)] + 1}{[f_1(t+1) - f_1(t)][f_2(t+1) - f_2(t)] + 1}$$

$$\begin{aligned} c_y = {} & \frac{1}{2}\frac{f_2(t+1)[f_1(t+1)^2 - f_1(t)^2 + 1]}{[f_2(t+1) - f_2(t)][f_1(t+1) - f_1(t)] + 1} \\ & - \frac{1}{2}\frac{f_2(t)[f_1(t+1)^2 - f_1(t)^2 - 1]}{[f_2(t+1) - f_2(t)][f_1(t+1) - f_1(t)] + 1} \end{aligned} \tag{2}$$

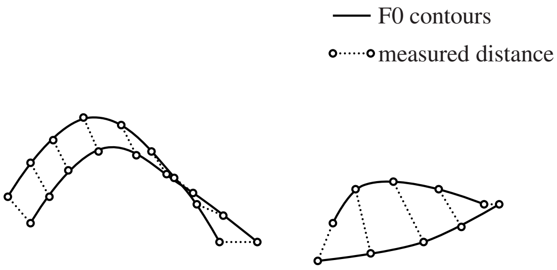

Figure 2: Illustrating the warp method.

and:

$$r_x = 0.5$$
$$r_y = \frac{1}{2}[f_1(t) + f_1(t+1)] \tag{3}$$

The measured distances are then combined in same way that the RMSE calculation is computed.

3.2. Warping Method

the motivation behind the warping method come from the fact that frames of an F0 contour are generally not considered to be of equal importance, due to the fact that contour consists of important pitch events and less important intervening sections.

The warping method is designed to work on small sections of contours, specifically pitch events, although it could be used more generally on arbitrary sections of contour.

A label file which specifies the important sections of the contours is used to compute the metric only on the portions of the contour which are considered important.

A small section on each contour is isolated, and its length is calculated by summing the appropriately weighted distances between frame points, accounting for both time and F0 components. Each contour is then divided into a fixed number of equal length segments; the distances between the contours at the boundaries of these segments is used in the calculation, which is computed as the RMSE. This is illustrated in Figure 2.

Both the tangental method and the warping method include a threshold filter which discards distance greater than a predefined maximum. In the results presented here that distance is 100 for both metrics.

4. ANALYSIS OF RESULTS

4.1. Perception Experiment

We first examine the perceptual data alone. A Friedman test was carried out to test the hypothesis that subjects can make useful distinctions between F0 differences, and that their scores are not just random.

The test is significant with $\chi^2 = 170.32$, $df = 23$, $p = .000$, suggesting that the subjects scores are not random.

This allows us to continue and compare the perceptual scores with the computed scores.

The raw scores for each participant in the experiment were first rescaled onto the interval [0,4] by mapping the maximum and minimum scores used by a subject to the interval endpoints, and then linearly rescaling the rest of the scores appropriately. After rescaling it can be assumed that each subject used the same scale for their judgments. This allows us to compute an average score for each contour pair. This score can then be compared with the computational methods discussed above.

4.2. Comparison between perceptual results and computational methods

Although the use of the label file was specifically intended for the warping method, for completeness, the use of the label file is extended to all methods

Pearson correlation coefficients between the computation methods and computed scores are shown in Table 1.

The correlation coefficients show that the tangental and warping methods are comparable with the default RMSE and Pearson correlation methods, but the RMSE method does have the edge both with and without a label file. The warping method with a label file produces better results than the default RMSE calculation though.

It is worth noticing that all of the correlation coefficients are low. The highest being $R = .6534$ which gives $R^2 = .4269$ which means that only 42 percent of the variance within the scores can be accounted for by the metric. This is possibly due to the nature of the synthesised data. The pairs of utterances that are presented to the subjects were designed to be on a smooth continuum with respect to their RMSE scores. However, their perceptual score do not form such a smooth continuum. They were judged either as very similar, or distinctly different which what seems to be random variation across listeners.

These correlation results show the amount of variance within the variable score which can be accounted for by individual independent variable. This poses the question can they account for more variance if combined? This question is easily answered using multiple regression techniques, and the answer is no. The use of more than one variable cannot account for more variation in the perceptual score than the best of those variables alone.

	whole file				pitch events only			
	rmse	corr.	tangent	warp	rmse	corr.	tangent	warp
perceptual	.6441	-.5497	.6150	.6003	.6534	–	.5878	.6499
score	p=.000	p=.005	p=.001	p=.002	p=.001	–	p=.003	p=.001

Table 1: Pearson correlation coefficients of F0 contour distance metrics.

5. CONCLUSIONS

Our results suggest that the subjects can distinguish between identical and different intonation patterns but could not reliably distinguish the more subtle differences between the non-identical intonation patterns that they were presented with.

This illustrates an important factor which needs to be considered when evaluating synthetic intonation by comparison to natural intonation: if the synthetic intonation is distinctly different from natural intonation, small variations in differences between contours will not be perceivable by listeners. That is to say, small improvements to localised F0 are perceptually overshadowed by gross errors in the intonation as a whole.

This identifies the need for a testing subjects with a continuum of perceptual difference in pairs of contours and not with what are possibly only the two end points of this perceptual continuum, even if they form a smooth continuum of RMSE scores. This suggests that we may need to look for a non-linear relationship between our automatic scoring metrics and perceptual scores.

There is an alternative hypothesis which could account for the results we see which we can not completely rule out. This is that the subjects can make the distinction, but we are unable to produce a comparable metric to show this. That is to say that RMSE and the other metrics are just unsuitable measures of perceived difference, but this seems unlikely.

We then conclude that the perceptual experiment throws important light onto the the whole subject of evaluating synthetic intonation, which we need to account for in further experiments. The alternative metrics are comparable with the RMSE and correlation methods, but unfortunately no better. There is, however, possible room for improvement in the finer tuning of the weighting between time and frequency, to produce metrics which account for more variance in perceptual scores than RMSE. We finally conclude that RMSE performance can be improved by considering the sections of the speech contour which relate to pitch events discarding the sections which we consider unimportant.

REFERENCES

[1] K. Dusterhoff and A. Black. Generating F_0 contours for speech synthesis using the Tilt intonation theory. In *Proceedings of ESCA Workshop on Intonation*, Athens, Greece, 1997.

[2] A.L. Edwards. *An introduction to linear regression and correlation*. W.H. Freeman and Company, second edition, 1984.

[3] D. J. Hermes. Measuring the perceptual similarity of pitch contours. *Journal of Speech, Language, and Hearing Research*, 41:73–82, February 1998.

[4] K. Ross. *Modeling of intonation for speech synthesis*. PhD thesis, Boston University, College of Engineering, 1994.

[5] J. 't Hart, R. Collier, and A. Cohen. *A perceptual study of intonation: An experimental phonetic approach to speech melody*. Cambridge University Press, 1990.