Modelling Pitch Accents for Concept-to-Speech Synthesis.

Robert A. J. Clark

CSTR, University of Edinburgh
robert@cstr.ed.ac.uk

ABSTRACT

This paper addresses the problem of generating a full range of appropriate intonation contours for concept-to-speech synthesis systems where there is a specific requirement to produce different semantic meanings and contrasts through intonation. We show improvements through the appropriate use of prosodic structure and by improving the way in which individual accent shapes are modelled.

1 INTRODUCTION

The training of intonation models generally requires a reasonable amount of speech from a single speaker. Unfortunately, suitable speech corpora usually have a skewed distribution of pitch events: ”H*” and ”L-L%” are particularly frequent; ”L*+H” and ”H-H%” are particularly infrequent. Models trained on such data tend to generate the more frequent accents well, and the less frequent accents very badly. This is problematic, as the less frequent accents are the ones employed to convey specific meanings (e.g. questioning, uncertainty).

We describe methods of improving the performance for the less frequent accents without the need for completely intonationally balanced data. A Linear Regression model for use with the Festival [1] speech synthesis system is trained incorporating parameters which describe a suitable representation of prosodic structure along with ToBI [2] descriptions of pitch events. We then discuss ways to re-estimate particular parameters after training, using either other appropriate data or theoretical ‘ideal targets’ where data is unavailable. This results in a better model which can accurately generate a full complement of ToBI pitch events.

1.1 LINEAR REGRESSION MODELS

Contour generation in Festival is traditionally carried out using three Linear regression (LR) models [3]. Linear regression models assume that a predicted variable (p) can be modelled as the sum of a set of weighted real-valued factors.

$$p = w_0 + w_1 f_1 + w_2 f_2 + w_3 f_3 + ... + w_n f_n \qquad (1)$$

The factors (f_i) represent parameterised properties of the data, and the weights (w_i) are trained, usually using a stepwise least squares technique.

Each of the three models predicts the f0 at a different point of a syllable (start, middle and end respectively). The factors incorporate information like the type of accent present, the position of phrase breaks, syllable stress and syllable position within the text. Each model considers this information for a five syllable window centred on the current syllable. This allows the pitch on syllables around an accented syllable to be affected by the presence of the accent, so that pitch movement is not restricted to occur on the syllable that is marked with the pitch event. For example the peak of an L+H* could occur in the syllable following the one the accent is assigned to.

The training of these models requires a large amount (at least an hour) of good quality speech from a single speaker. This means that these models are inevitably trained on broadcast news. The phrase structure and distribution of pitch events found in broadcast news is somewhat idiosyncratic and not necessarily suitable for synthesis in other domains – or particularly good for building good intonation models from.

1.2 PROSODIC STRUCTURE

Broadcast news tends to comprise of long sentences each consisting of a number of prosodic sub-phrases. We exploit this by using the results [4] from an analysis of this structure to introduce parameters into the Linear Regression models enabling us to better account for sub-sentence level pitch range effects, which in turn helps us to model the pitch variation attributable to accents and boundaries more accurately.

We take the ToBI break indices assigned to the data as our starting point and derive two levels of phrasing. We will call our two levels of phrasing *IP* and *TG*. These terms are loosely based on those of [5, ch. 6] but are not necessarily meant to relate to phrase units of the the exact same size and type. An IP may in some circumstances be thought of as an *intonation phrase* but we do not wish to call it that explicitly because it may or may not directly relate to what others, particularly [6] call an intonational phrase. Similarly a TG can be thought of as a *tone group* which we consider to be a sequence of tones ending in some kind of boundary, and nothing more. An *utterance* then consists of one or more IPs each of which in turn consists of one or more TGs. We classify TGs with a three way initial/medial/final distinction based on the findings of [7]. This representation of structure is then incorporated into the linear regression models by including parameters representing TG type.

2 THE MODELS

2.1 THE INITIAL MODEL

We initially build three linear regression models which incorporate parameters representing the TG structure described above along with parameters representing accent type, accent position and other standard parameters which account for a syllables position within the TG (see [7] for full details). We have two versions of this model, one which works with a full complement of ToBI labels and one which works with a more general set of accent descriptions: ‘a’ for accent, and ‘fb’ and ‘rb’ for falling and rising boundaries respectively.

As expected, the resulting model is found to generate good contours for unaccented parts of utterances and for H* accents and L-L% boundaries. Other less frequent accents and boundaries, L* L+H*, L*+H, L-H% and H-H% result in less acceptable contours. Closer examination of the training data shows that there is a large amount of variation in the contour shape for these pitch events, and they occur much less frequently than the H* and L-L% events which the model captures well. The models using the more general accent descriptions generally perform better, but only because an ‘a’ accent can usually be considered to be an H* accent, and the other accent types are not dealt with by this model. With this in mind, we consider ways to re-estimate the parameters which are specific to these minor accents to improve the performance of this model

2.2 THE ENHANCED MODEL

We find that the mutual exclusivity of the parameters which account for different accent types allow these parameters to be re-estimated without affecting the rest of the model. Re-estimation using other ToBI labelled data which better represents the pitch events in question was first used to retrain these parameters. However, the problems associated with the arbitrary re-scalings that were required to map the pitch ranges of these other speakers to match the pitch range of the original speaker, made it clear that defining an ‘idealised’ accent shape was a better approach than trying to use real data that was inconsistent with the original speaker.

2.3 THE RE-ESTIMATION PROCEDURE

The ability to re-estimate certain parameters is based around the fact that the model is designed in such a way so that the parameters that control accent shape either make a contribution to pitch that is independent of other parameters, or they are the leaves in a hierarchical dependence structure.

The parameters are adjusted by first synthesising a short example utterance using the accent or boundary for which the parameters are to be changed. The resulting contour is then compared to an ‘idealised target’ contour which we want the model to generate. Recall that that our models predict f0 target points at three points in each syllable and over a five syllable window. For our purposes here we only modify the parameters relating to a three syllable window—the accented syllable and those either side of it—we often only need a two syllable window and do not modify the syllable before the accented syllable.

For each point in our window we calculate the error between the generated contour and our idealised example, then we adjust the original parameter value by this error. Table 1 shows original and adjusted weights for the the L* accent and figure 1 shows the difference in the resulting pitch contour.

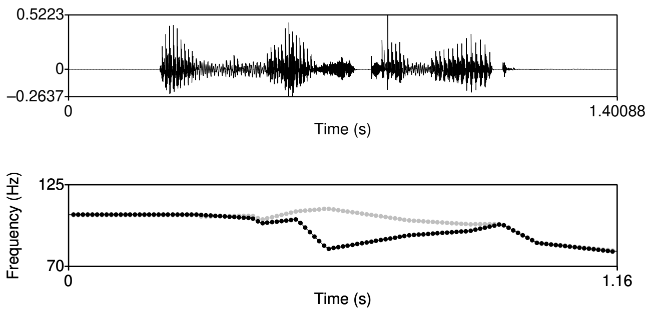

Figure 1: Example L* contour

	accented syllable			next syllable		
	s	m	e	s	m	e
Original Model	0	-27	-26	-14	0	0
Adjusted Parameter	-5	-66	-43	-14	0	0

Table 1: Parameter adjustments for L*

3 EVALUATION OF THE MODELS

Two simple perceptual evaluation experiments were performed, where listeners were asked to judge which of a pair of utterances was ‘most appropriate’. The intention was to show that the intonation produced by the models is reasonable, based on a qualitative measure. The first hypothesis we test is that subjects prefer natural intonation over synthetic varieties, but prefer the above models over earlier synthetic models. The second hypothesis tested was to show the enhanced model was preferred over the initial model.

To try to control for segmental quality all the speech was created by diphone synthesis using Festival. The overall speech rate was also controlled as much as possible to provide comparable utterances. Diphone resynthesis with natural segment durations and pitch contours was used to create the natural intonation patterns.

3.1 EXPERIMENT I

To test the first hypothesis, subjects were presented with three example sentences: A short sentence, and a longer sentences from the broadcast news domain that the model is trained upon, and a longer out-of-domain sentence describing a museum exhibit. Each sentence pair was presented ten times, with the order of presentation swapped for half of the presentations. Utterance pairs were presented in a random order to each subject.

To judge appropriateness the listeners were asked to judge which of each pair they thought most appropriate for a given style of speech: the broadcast news style, in the case of the short sentence and long sentence from this domain, and a style suitable for a museum guide for the other sentence.

Twenty seven native English speaking subjects took part in the experiment. As we suspected, subjects found it difficult to consistently make the kind of judgements we are asking for. So we excluded subjects if their consistency in judging repetitions of stimulus pairs fell below 95%. Thirteen out of the twenty seven subjects were found to be consistent, and the results of these subjects were analysed further. Figure 2 shows the distribution of consistent and inconsistent judgements. The thirteen consistent subject are clearly seen at the right tail of the distribution that would be seen if the subjects were making random judgements.

Table 2 shows the preferences of the the consistent subjects. All results except the 50/50 result are significant at $p < 0.05$. There is a surprising aversion to the natural intonation for the short sentence and the long broadcast news sentence, but a clear preference for the natural intonation for the museum object description. Our new model

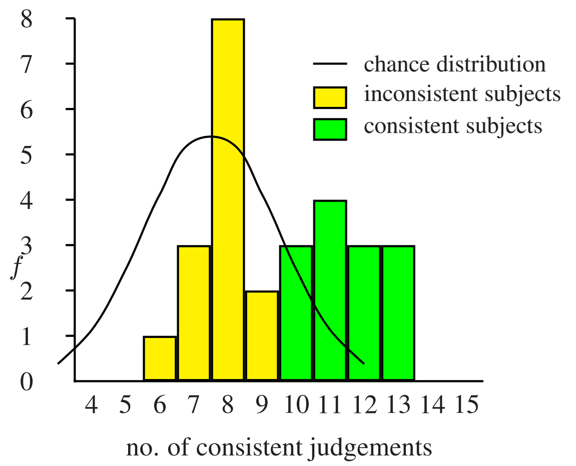

Figure 2: Frequencies of number of stimuli pairs judged as consistent by individual subjects.

is preferred over our baseline for the short sentence and for the museum description. There was however no preference between the models for the long broadcast news sentence. Our belief is that subjects find it difficult to make consistent judgements for longer sentences.

	Natural vs Model	Model vs Baseline
Sentence		
Short	81 / 19	77 / 23
Long	73 / 27	50 / 50
Museum	8 / 92	65 / 35

Table 2: Consistent speakers preferred models

3.2 EXPERIMENT II

To test our second hypothesis, than our enhanced model is better that our initial model, a second experiment was carried out using the same basic methodology as used for the previous experiment. The sentences used in this experiment were taken from the MagiCster project [8]. The text used was an example paragraph of a doctor giving a patient a diagnosis. This text is the output of a language generation system and is marked-up with appropriate prosodic phrasing and pitch accents with the intention of conveying a particular meaning.

Five relatively short sentences were taken from the paragraph of diagnosis and synthesised using both the initial model and the enhanced model.

Each sentence pair was presented 4 times to each subject, twice with the initial variant presented first and twice with the enhanced variant presented first. The 4 variants for each sentence were then combined making 20 stimulus pairs in total. These were then presented to the subjects as a single block in a different random order for each subject.

Six of the subjects who were found to be consistent in the previous experiment were chosen to take part in this experiment giving a total of 120 responses, 24 for each sentence. The subjects were asked to decide which of each pair they thought sounded the most natural.

The total responses for each sentence are shown in table 3. Overall 79 out of the 120 stimuli pairs presented showed a preference for the enhanced model over the initial model. This is significant at $p < 0.01$ in a binomial test. Looking at the results sentence by sentence it is clear that the enhanced model is preferred for sentences 1, 3, 4 and 5, but the initial model is preferred for sentence 3. Each of these individual results are significant at $p < 0.05$.

The level of consistency in this experiment was lower than that found in the previous experiment. This was to be expected as the difference between stimulus pairs here was much less than in the previous experiment, as the models being used only produce localised difference in pitch around particular pitch events. We interpret the low level of consistency as meaning that the subjects found this task particularly difficult. This interpretation was reinforced by subjects comments after the experiment saying that it was harder than the previous experiment.

The discrepancy in the results relating to sentence 3 is thought to be due to a rising boundary specified in the mark-up in a place where the it sounds more natural without one. As the initial model fails to generate a convincing rise it is judged better than the enhanced model which does generate it, reminding us that a model is only as good as its input.

4 CONCLUSIONS

We have shown that we can improve the intonation for speech synthesis in two ways, firstly through the use of appropriate prosodic structure and phrasing, and secondly through the re-estimation of parameters to account for deficiencies in the data available to train models on. We have shown through perceptual experiments that these improvements can be recognised by listeners and that they are statistically significant.

ACKNOWLEDGEMENTS

The author wishes to acknowledge the EC projects: MagiCster (Embodied Believable Agents IST-1999-29078) and M-PIRO (Multilingual Personalised Information objects IST-1999-10982) for their funding and support of this research.

	Initial	Enhanced
Sentence 1	6	18
Sentence 2	6	18
Sentence 3	17	7
Sentence 4	6	18
Sentence 5	6	18
Total	41	79

Table 3: Distribution of responses for experiment evaluating the naturalness of the enhanced model.

REFERENCES

[1] A. Black, P. Taylor, and R. Caley, "The Festival speech synthesis system," http://festvox.org/festival, 1998.

[2] Kim Silverman, Mary Beckman, John Pitrelli, Mari Ostendorf, Colin Wightman, Patti Price, Janet Pierrehumbert, and Julia Hirschberg, "ToBI: A standard for labeling English prosody," in *Proceedings of the 1992 International Conference on Spoken Language Processing*, 1992, pp. 867–870.

[3] A. Black and A. Hunt, "Generating f0 contours from ToBI labels using linear regression," in *ICSLP 96*, Philadelphia, Penn., 1996.

[4] Robert A. J. Clark, "Using prosodic structure to improve pitch range variation in text to speech synthesis," in *XIVth International Congress of Phonetic Sciences*, 1999, vol. 1, pp. 69–72.

[5] D. Robert Ladd, *Inonational Phonology*, 1996.

[6] Janet Pierrehumbert and Julia Hirschberg, "The meaning of intonational contours in the interpretation of discourse," in *Intentions in Communication*, P. R. Cohen, J. Morgan, and M. E. Pollack, Eds., chapter 14, pp. 271–311. 1990.

[7] Robert A. J. Clark, *Generating Synthetic Pitch Contours Using Prosodic Structure*, Ph.D. thesis, University of Edinburgh, 2003.

[8] "Embodied believable agents," IST-1999-29078, 2002, http://www.ltg.ed.ac.uk/magicster.