Generating F_0 contours from ToBI labels using linear regression

Alan W Black[1] *Andrew J Hunt*[2]

ATR Interpreting Telecommunications Laboratories,
2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-02, JAPAN.
awb@cstr.ed.ac.uk, hunt@winston.east.sun.com

ABSTRACT

This paper describes a method for generating F_0 contours from ToBI labelled utterances. The method uses linear regression to predict F_0 target values for the start, mid-vowel and end of every syllable, using features representing the ToBI labels, stress and syllable position. Contours generated by this method for an English database have a correlation of 0.62 and 34.8 Hz RMS error when compared with originals from test data. These results are significant improvements on a previous rule driven method (0.40 and 44.7), and the new method contours are preferred by human listeners. The technique has also been successfully applied to Japanese ToBI with similar improvements.

1. INTRODUCTION

One problem in the process of synthesizing natural sounding speech is the prediction of an F_0 contour which adequately reflects the desired prosodic tune. In most synthesizers the task of generating a prosodic tune consists of two sub-tasks, the prediction of intonation labels (accents, tones, etc) from text and the generation of a contour from those labels (and possibly other information). This paper deals solely with the second of those tasks.

The experiments presented here look at one particular intonation phonological labelling system and improve on an existing method of generating an F_0 contour from these labels. The ToBI labelling system [7] offers a method for labelling pertinent aspects of intonation in speech. Although there are recognized limitations with the system, it has been used to hand-label large speech databases and is being used in a number of synthesis systems.

This work has been fully implemented in ATR's CHATR speech synthesis system [2], thus showing the new technique not only produces better F_0 contours from hand-labelled natural utterances but also for fully synthesized utterances.

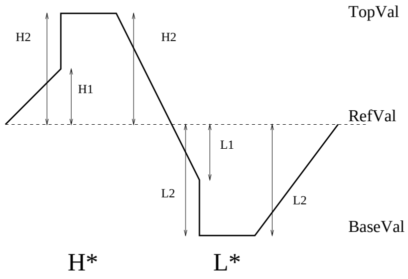
2. RULE-DRIVEN METHOD

The ToBI labelling system has its origins in the Pierrehumbert intonation labelling system [6]. ToBI in fact has been interpreted in a number of slightly varying ways but could be generally defined as follows. A ToBI labelling for an utterance consists of three tiers each related (through time) to a speech waveform. The tiers are: labels, breaks indices and miscellaneous. The label tier marks *pitch accents*, *phrase accents* and *boundary tones*. The break index tier marks one of four levels of prosodic breaks. The miscellaneous tier may contain any other labelling, such as background noise, coughing, laughing, disfluencies or anything else that might be labelled.

From the synthesis point of view, the question is how well can we predict a F_0 contour using the labels and breaks. Although useful information may exist in the miscellaneous layer, it is not formally defined what exists (and what does not) therefore it is ignored.

One method for generating an F_0 contour from such labels and breaks is described in [1], which we will call the *APL method.* A similar generation method for a Japanese version of ToBI is described in more detail in [5, chap7]. We will briefly describe the APL method as it is this we wish to improve on.

The APL method predicts a number of *target points* for each syllable marked with a pitch accent, phrase accent or boundary tone. A number of specific rules deal with each case. For example consider the following diagram



An `H*` accent introduces three target points, the first of

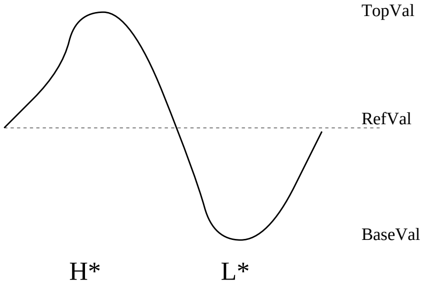
[1] Now with Centre for Speech Technology Research, University of Edinburgh, Scotland.

[2] Now with Sun Microsystems Laboratories, Chelmsford, MA.

height H1 above the reference line at the start of the syllable, the second at height H2 also at the start and the third also at H2 at the end of the syllable. Similarly for L* except they are below the reference line.

The parameters H1, H2 etc. are given as fractions of TopVal and BaseVal above or below RefVal, so there is some independence from absolute pitch range. Independently RefVal TopVal and BaseVal may decrease over time to represent declination.

After all targets for labelled syllables are predicted they are smoothed to produce a more familiar F_0 looking something like



Special rules are required for syllables labelled with multiple labels, such as accents and ending tones, causing the targets to be squeezed appropriately.

The various parameters may be set by hand and experimentation (though some experiments to extract these values from data are described in [5, chap7]). For most implementations these values are set by hand and tuned until acceptable results are achieved.

3. LINEAR REGRESSION METHOD

Instead of trying to adjust the parameters for the model described above we are interested in automatically finding the optimal values of these parameters. Although some form of gradient descent algorithm could be used for optimization, a more simple quicker approach was undertaken. The approach was simply to predict three F_0 target values for every syllable, one at the start of the syllable, one in mid-vowel and one at the end of the syllable. Prediction uses the formula

$$target = I + w_1 f_1 + w_2 f_2 + w_2 f_2 + \ldots + w_n f_n$$

Where f_i are features that are felt to contribute to the F_0 value, such as accent type, position in phrase etc. I and the weights w_{1-n} are estimated from data using linear regression.

Accents (i.e. ToBI pitch accents) are represented by 5 binary features, each representing the group an accent falls within. The accent groups are: accent_1: H*, accent_2: !H*, accent_3: L*, accent_4: L+H* L+!H* H+!H* L*+!H L*+H, accent_5: other. The complex accents are grouped because of their frequency within our database is low, if more data were available these could easily be split.

Phrase accents (H- and L-) and boundary tones are grouped together as the grammar of ToBI does not allow them to co-occur. They are grouped into 6 classes: endtone_1: H-, endtone_2: L-, endtone_3: L-L%, endtone_4: L-H%, endtone_5: H-L%, endtone_6: other.

The third ToBI related set of features encodes break indices. Instead of encoding the break index as a single value we encode it as 4 separate binary features, depending on which break index (1 to 4) this allows a certain amount of non-linearity.

Thus for each syllable we collected the following features

the accent type on this syllable and that of the previous two syllables and following two syllables.
the endtone type on this syllable and the previous two syllables and following two syllables.
the break index type on this syllable and the previous two syllables and following two syllables.
the lexical stress of this syllable, and the two previous and two following syllables.
the number of syllables from start and to end of current phrase
the number of stressed syllables from start and to end of current phrase
the number of accented syllables from start and to end of current phrase
the number of syllable since last accented syllable (and to next accent)

Using the same set of features for each syllable we build three linear regression models predicting the start F_0, mid-vowel F_0 and end F_0 respectively.

In generation the predicted targets are smoothed and interpolated to give a continuous contour which is applied to the waveform using PSOLA [3].

4. COMPARATIVE RESULTS

The above model was tested on the Boston University FM Radio corpus [4] for speaker f2b. F2b consists of about 45 minutes of female American news reading speech. It has been hand-labelled with ToBI labels, though the documentation admits there may be some inconsistencies in the labelling. The data used in this experiment consists of 14,778 syllables around 67% of which are unaccented, 20% H*, 5% !H*, 4% L+H*, 1.6% L+!H*, 1.3% and others 1%.

In predicting an F_0 target value it is necessary to decide what such a value might be during unvoiced segments. As it is the *contour* we are trying to predict it was decided to use an interpolated contour over the whole utterance (except during pauses). This was done for primarily three reasons: first as a full contour is, in our system, presented to our prosodic modification module (PSOLA based), second smoothing a set of target points where some are forced to zero for phonetic reasons would be complex, and third we did not wish to include phonetic properties in our contour prediction which would require significantly more data.

For our training data we first extracted a raw contour using a standard pitch tracker (ESPS's `get_f0`). Using the F_0 values and the vocing information we constructed our *smoothed contour* by finding the mean F_0 for voiced sections of all segments and interpolating between them (ignoring any segment with no voiced sections at all). A 10 ms frame by frame comparison between the smoothed contour and raw contour (ignoring frames marked unvoiced in the raw contour) offers an RMS error of 9.9 Hz with a correlation of 0.90. Note that time alignment is not a concern. As original durations are used throughout, both contours will always be aligned.

From this training data we extracted the start, mid-vowel and end values for each syllable. We split out data into training and test data (12000/2778) and built three linear regression models for the training data. For the individual models we achieved the following results

	Train		Test	
	RMS	Corr	RMS	Corr
start	27.1	0.53	27.4	0.55
mid-v	26.2	0.66	26.1	0.68
end	27.7	0.56	28.4	0.55

In order to find out if these results are an improvement we wish to compare them with the contours produced by the APL method described above. We cannot directly compare them as the APL method does not explicitly predict start, mid-vowel and end target points therefore we took the LR model's results and interpolated between them and compared the continuous contours as predicted by the LR method and the APL method with the smoothed F_0 we used for the LR model training. The results were

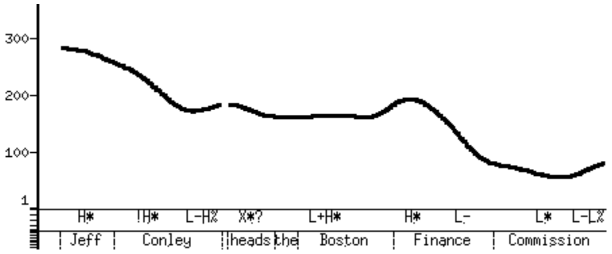
	RMS	Corr
APL	44.7	0.40
LR	34.8	0.62

These figures suggest that the LR method is producing a contour much closer to our smoothed original.

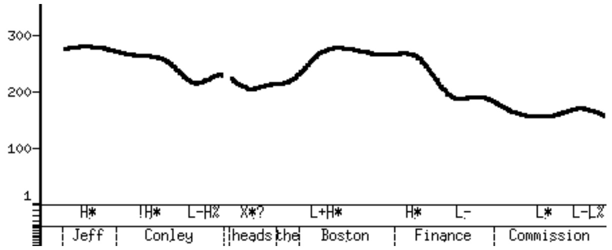
As a further test, we implemented this algorithm within CHATR and synthesized a number of utterances from the test set using the APL and LR models. In each case we took the original "natural" information from the utterance (i.e. segments, durations, ToBI labels etc.) and used it to predict the F_0 contour, Then we used PSOLA to impose this contour on the original natural utterance. 10 (short) sentences from the test set were chosen and played to three native English speakers (not including the authors). In 70% of the cases listeners preferred the LR F_0 contours over the APL generated ones.

The following examples offer a comparison between the techniques, and the original. Because our PSOLA implementation introduces distortion, we include the original similarly distorted to offer a fair comparison. Note the waveform example contains the full sentence while the graphs, for ease of reading only contain the first clause *"Jeff Conley heads the Boston Finance Commission,"*

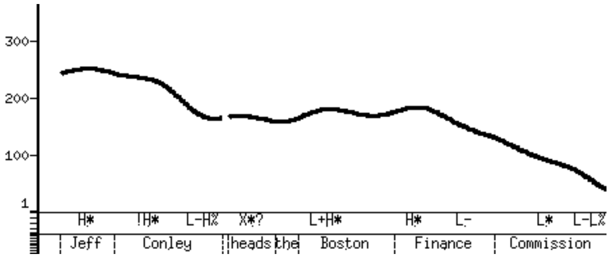
The following graph shows the original smoothed F_0 [SOUND A803S01.WAV]



The contour generated by APL method for the same utterance shows a much more varied contour [SOUND A803S02.WAV]



While the LR method produces [SOUND A803S03.WAV]



In general LR produced F_0 contours were less varied than those generated by APL, sometimes causing them to sound not as "interesting" as the APL counterpart. However the APL generated contours often sounded over-varied and inappropriate (something the LR ones never did). A more detailed critical comparison is included below

5. JAPANESE TOBI

To further test this LR model we applied it to a Japanese databases marked with Japanese ToBI (JToBI) as described in [5]. Using the same technique we used a database of 503 sentences spoken by a male (Tokyo) Japanese speaker (ATR MHT Bset). We compared the result against a previously existing implementation of that described in [5] which although caters for the different intonational phonology (i.e. for Japanese) is effectively similar to the APL technique for English described above. The following results were achieved for full generated contours when compared with the smoothed original.

	RMS	Corr
APL	25.6	0.55
LR	20.9	0.70

The results were played to a Japanese native speaker and although in general expressed a preference for the LR sentences, the APL sentences were never particularly bad. The better results for Japanese are probably due to a number of simple reasons. First Japanese intonation is probably less varied than English, the Japanese speaker (MHT) is very consistent, MHT's pitch range is less than f2b's. Also more time was spent tuning the Japanese APL parameters than for English. But it should be noted that the Japanese parameters were tuned over a year while the LR model was trained and applied in an afternoon. It gives better results and is immediately customizable to other speakers.

6. DISCUSSION

From the above results it appears that the linear regression method better models an F_0 contour. It is fully trainable and shows improvement over previous techniques even for multiple languages. The implementation of the LR method is substantially simpler than the APL method. The APL method requires devising particular shapes for the various ToBI labels and creating parameters to define the size and position of these shapes. Also the APL method includes a separate feature for declination. All the parameters used to realise these features need to be given values, so far in our implementations, by hand, although some training method could be devised. In the LR case it is simply a matter of collecting the feature values for each feature and summing their weighted values, where the weights are explicitly available from the training method.

When a ToBI labelled database is not available for training the results from a different speaker, of the same dialect, may be transfered. Absolute F_0 target values from the trained model are converted into zscores (i.e. number of standard deviations from the F_0 mean). Those zscores may be converted into the target speaker's range using the F_0 mean and standard deviation of the target speaker. This technique (which can also be used for the APL model) has proved quite adequate.

However in spite of the advantages there are distinct disadvantages of this technique too. There is no way this technique will learn contours for labels not in the training database, or labels with few examples. Particularly, in the English case, the f2b news database contains only three `H-H%` boundary tones. This is insufficient for the model to learn about final rises and hence when presented with a syllable marked with `H-H%` the resulting F_0 does not rise. This problem does not exist when using the APL method where explicit rules for each label are devised. A second problem is in syllables with multiple labels, where the intonation contour cannot necessarily be captured by three target points alone. Such phenomena are rare in news speech though are more common in say dialogue speech.

Although we can assume that in databases of dialogue speech there will be more examples of the labels such as `H-H%`, and likewise databases will have labels representing their intonational variation, the above LR method may in fact be too general. A more specific model may give better results, (especially in cases where there are only a few examples of particular labels). We can look at the APL model and LR model on a scale. The APL model requires specific rules for each accent (and combination) that can exist on a syllable. The LR model however treats all syllables in the same way irrespective of their labels, i.e. effectively a single rule. Some medium may be better where specific label types (or clusters of labels) have specific patterns of targets. As yet this space of possible methods has not yet been investigated.

In conclusion, the results show that the linear regression method presented here offers a better modelling of the F_0 contour than previous published method, also the model does not require special rules for each label type. The model is general enough for both English and Japanese.

7. REFERENCES

1. M. Anderson, J. Pierrehumbert, and M. Liberman. Synthesis by rule of English intonation patterns. In *Proceedings of ICASSP 84*, pages 2.8.1–2.8.4, 1984.
2. A. W. Black and P. Taylor. CHATR: a generic speech synthesis system. In *Proceedings of COLING-94*, volume II, pages 983–986, Kyoto, Japan, 1994.
3. Eric. Moulines and Charpentier F. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, 9(5/6):453–467, 1990.
4. M. Ostendorf, P. Price, and S. Shattuck-Hufnagel. The Boston University Radio News Corpus. Technical Report ECS-95-001, Electrical, Computer and Systems Engineering Department, Boston University, Boston, MA, 1995.
5. J. Pierrehumbert and M. Beckman. *Japanese Tone Structure.* The MIT Press, Cambridge, Mass., 1988.
6. Janet B. Pierrehumbert. *The Phonology and Phonetics of English Intonation.* PhD thesis, MIT, 1980. Published by Indiana University Linguistics Club.
7. K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg. ToBI: a standard for labelling English prosody. In *Proceedings of ICSLP92*, volume 2, pages 867–870, 1992.