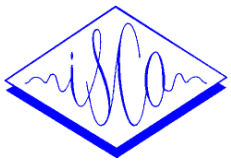

3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

COMBINING THE USE OF DURATION AND F0 IN AN AUTOMATIC ANALYSIS OF DIALOGUE PROSODY

Nick Campbell

ATR Interpreting Telecommunications Research Laboratories,
2-2 Hikaridai, Seika-cho, Kyoto 619-02, Japan. nick@itl.atr.co.jp

ABSTRACT

This paper evaluates the performance of two automatic labelling systems for intonation by using their output to predict the position and type of prominences and boundaries labelled in a ToBI transcription of twelve dialogues. It shows that they both model the prosodic information well, and that improvements in modelling are gained when including segmental information about the duration and energy profiles of the utterance. However, after parameter reduction, the features that survive are segmental rather than f0-related.

1. INTRODUCTION

With the recent growth in the number and size of speech corpora, there is an increasing need for automatic systems to label the data collected, both segmentally and in terms of the prosody of the speech. Previous studies on the fundamental frequency and segmental lengthening in English speech have shown that there are strong relationships between these acoustic prosodic parameters and corresponding linguistic prominences and boundaries in the utterance. Wang & Hirchberg '92 [9] described a system for the automatic classification of intonational phrase boundaries using binary decision trees with syntactic and other information derived from the text. However, one of their test criteria requires observed boundary information (for calculation of elapsed phrase length), which prevents use of the model in a fully automatic system. Wightman & Ostendorf '94 [10] also propose a model for the automatic prosodic annotation of speech, based on tree clustering, which recognises boundary location and strength, and degree of prominence, but not accent or tone type.

This paper builds on previous work in an attempt to combine and utilise only acoustic and segmental sources of information such as would be available from a fully automatic analysis of the speech waveform. Taylor '94 [8] proposed an f0 model that labels pitch accents and boundary tones, using only fundamental frequency contours as input. This cannot yet be done automatically, and requires manual intervention to pass from the phonetic to the phonological levels, but a pre-processing stage produces an abstraction of the f0 that might be adequate for contour stylisation. Hirst '93 also describes a model for the stylisation of f0 that is fully automatic. Campbell '93 [1] describes a model taking normalised segmental durations as input that distinguishes the lengthening caused by prominence from that arising from phrase-final, pre-boundary position.

The weakness of both the pitch modelling and duration-based systems is that neither takes into account the corroborating information from the other. Neither system in isolation is able to reliably detect all the major prosodic events in the speech data. In the case of segmental durations, there are frequently segments that are lengthened from both prominence and pre-boundary position, and in some instances the prominence or boundary events are realised by other factors than lengthening. Similarly, there are accents that cannot be detected from analysis of the intonation contour alone; for example, a series of downstepping tones may be difficult to distinguish in a falling contour. A combination of the two information sources, intonational and segmental, may offer a more robust, unified prosodic description.

2. MODELS OF THE PITCH CONTOUR

Two models of pitch contour stylisation were examined in this study. The first, a rise-fall-connection model (*rfc*) (Taylor '92 [7], '94 [8]) encodes a heavily smoothed fundamental-frequency contour into a series of rise and fall elements, with connection elements linking the events in the signal that are marked by the rising and falling pitch. The second, a target model (*tgt*) (Hirst '80 [3], '93 [4] implicitly factors out the microprosodic elements of the contour and encodes the macroprosodic element as a series of turning or target points.

Both models use quadratic spline functions (that interpolate linearly between points of which the first derivative is zero) to model the curvature of the contour and extract symbolic representations. They differ primarily in that one puts out a description of the turning points, and the other puts out a description of the lines between them. Both include local contour gradient information in their encodings, but Hirst's gives more information about height, while Taylor's gives slope and rate of change.

This paper examines the degree of correlation of each with a hand-generated transcription of a set of test utterances, and shows that a better fit can be obtained if we also take into account non-pitch information about lengthening and energy changes in the speech. No attempt is made here to optimise the prediction of the joint model, and the purpose of this study is simply to determine if it would be advantageous to combine the information sources and, if so, which f0 contour abstraction system is to be preferred.

3. A MODEL OF NON-PITCH CONTOURS

Campbell '93 [1] described a method (*wnc*) for automatic annotation of prosodic events in speech using segmental duration information alone. This was derived and tested using phonetically-balanced sentences, and was not tested on natural dialogue.

In this model, raw data in the form of segment labels and their durations are processed to provide a normalised measure of the lengthening undergone by each segment (z-score normalisation per phone class). A gradient of lengthening within the syllable (the difference between syllable onset and rhyme) is used to distinguish lengthening due to prominence or stress from lengthening due to proximity to a phrasal boundary.

$$length_{seg} = \frac{(observed\ duration_{seg} - mean_{phone})}{standard\ deviation_{phone}} \quad (1)$$

A further acoustic measure is here included. For each segment in the speech, as marked by a label in the input, raw measures of rms energy were extracted from the waveform, averaged across the segment duration, and normalised by phone class as described above for the durations.

In addition to these absolute measures of relative lengthening and energy change, a further transform is applied to model their contours and, to encode the fact that a given segment is part of a rising or falling (lengthening or shortening) part of the utterance. To do this, first derivatives of the normalised scores were calculated on averages taken across three phones previous and following for each segment; the difference of these averages being taken as a measure of the local contour shape in the vicinity of the segment. The sign of the result indicates whether a segment is part of a rising contour (increasing loudness or length) or falling; its magnitude indicates the rapidity of the change.

$$contour_t = \sum_{i=1}^{3} segdata_{t-i}/3 - \sum_{i=1}^{3} segdata_{t+i}/3 \quad (2)$$

For each syllable in the test database and for each measure (excursion from the mean, and local contour shape), the maximum, minimum, and mean across the syllable were calculated along with the raw values for the last segment in the syllable. Each utterance was classified in terms of these energy and lengthening values and comparisons were made with hand-labelled prominences and boundaries, both individually, and in conjunction with the two pitch contour models.

4. COMPARING THE MODELS

The purpose of this prosodic labelling is to help parse an utterance in the context of speech understanding for an interpreting telecommunications application, and to provide large prosodically transcribed databases for speech synthesis research in the same application environment. It is desirable that we be able to a) distinguish phrasing, b) detect the salient words within each phrase, and c) determine the relationship between the phrases. Since we are translating dialogue, it is also helpful to extract speech-act and discourse cues for turn-taking and for the type of response required.

4.1. A baseline for comparison

Data for the test consisted of twelve dialogues from the CMU-ATR conference-registration database. This is not 'natural speech' as it was collected from students who were paid to perform for the recordings, but is closer to spontaneous dialogue than to read speech. The data used are from one male speaker (*maem*) who took both roles in the conversation, simulating both receptionist and guest, to provide a complete set of dialogues. Recordings were made under noise-free conditions, and the speech data sampled at 16kHz. The twelve dialogues ranged in length from 11-32 turns (206 in total) and included 1732 words. Turns ranged in length from 1 to 118 words, with an average of 19 words per turn. Silences within an utterance were treated as ordinary syllables.

In order to have an objective basis for evaluation of the automatic pitch analysis models, the twelve dialogues were labelled according to the ToBI system of prosodic annotation [6]. This system marks each word with a break index, each accented syllable with a pitch accent type, and each phrase with a phrasal tone type. The break index shows the degree of separation from the following word, and can also be taken as an indication of the phrasing of the utterance. The pitch accent indicates salient syllables and distinguishes five types of peak accent; if a syllable is not accented, it will not be marked. The phrasal tone marks two types of prosodic boundary (major and minor) and is of importance in the interpretation of the discoursal role of the utterance; it can help distinguish questions from statements, and provides cues to whether the speaker intends to continue with the same utterance, or is ready to be interrupted. The ToBI labelling produced 1,093 accented and 1,620 unaccented syllables. Table 1 shows counts for the ToBI labels from the 12 dialogues.

pitch accent		phrasal tone		break index	
H*	846	H-	56	0	37
H+!H*	77	H-H%	19	1	881
L*	55	H-L%	51	2	326
L+H*	115	L-	56	3	125
unaccented	1620	L-H%	68	4	295
		L-L%	159		

Table 1 ToBI label counts for the 12 dialogues.

While it has not yet been confirmed that this type of labelling can be of use in automatic language understanding, it is clear that speakers do use prosody to aid the interpretation of an utterance, and it has also been shown that labellers can produce consistent transcriptions under the ToBI notation scheme. Silverman *et.al.*, have reported approximately 85% agreement on phrase accent type (90% agreement on position), 78% agreement on pitch accent type (83% agreement on presence or absence of an accent), and 94% agreement on break index labelling (within ± 1). Later results of a more extensive test are to appear elsewhere in these proceedings. It is by no means certain that a ToBI labelling of such a database will provide a useful criterion for evaluation, but it does at least provide a baseline against which to compare the results of the automatic systems in an objective manner.

4.2. Auto labelling with tgt and rfc

The pitch modelling programs were applied to f0 contours from the speech waveforms extracted with *get_f0* (esps/waves+ [2]) and their output labels were compared on a syllable-by-syllable basis with the ToBI labels. The parameters of the *rfc* model had been tuned to this database, those of the *tgt* model were set to the default for a male speaker. Means and standard deviations used in the *wnc* model were derived from the target database. All models were run in fully automatic mode, and no manual correction was performed.

x	:	880	hl	:	38
l	:	425	mh	:	35
h	:	407	uh	:	29
d	:	232	dl	:	28
u	:	150	mhl	:	24
e	:	149	hd	:	23
mu	:	66	he	:	19
lh	:	50	lu	:	19
m	:	46	muh	:	15

other: 33 classes (max len = 5)

Table 2. tgt label counts for the 12 dialogues.

The output from *tgt* was in the form of time locations and target points. Following Hirst's convention (*pers. comm.*, '94) these were interpreted to produce a set of labels (U: up, D: down, H: high, L: low) showing target type, and a further two labels (M: start, E: end) marking beginnings and ends of the utterance. In the cases where more than one target was determined within a syllable, the target values were averaged and the labels concatenated. Counts are given in Table 2.

The output from *rfc* was in the form of time locations with values for start-time and duration of the label, its starting f0, and the amount of rise or fall. As above, labels were concatenated when more than one was marked on a syllable, and the values for rise and fall (in f0) were averaged.

```
x  :     1073      | rf  :      73
c  :      444      | cf  :      55
f  :      381      | crf :      31
r  :      302      | fr  :      25
fc :      118      | fcr :      20
cr :      137      | rfc :      12
  other: 15 classes (max len = 4)
```

Table 3. rfc label counts for the 12 dialogues.

4.3. Growing a classification tree

The assumption underlying the test was that each of the models would produce a set of labels that in some way captured the significant events in the prosody of the utterances, and that these label sets could be correlated with the hand transcribed labels and used to determine prominences and boundaries in the data.

```
dz < 2.11:
|   ez < -1.595: [0.32 0.68] Yes
|   ez >= -1.595:
|   |   dz < 1.465:
|   |   |   dz < 0.685: [0.95 0.05] No
|   |   |   dz >= 0.685:
|   |   |   |   ez < -1.175: [0.11 0.89] Yes
|   |   |   |   ez >= -1.175:
|   |   |   |   |   dzm < 0.535: [0.81 0.19] No
|   |   |   |   |   dzm >= 0.535:
|   |   |   |   |   | edm < -0.31: [0.99 0.01] No
|   |   |   |   |   | edm >= -0.31: [0.33 0.67] Yes
|   |   dz >= 1.465:
|   |   |   edm < -0.555: [0.99 0.01] No
|   |   |   edm >= -0.555:
|   |   |   |   ez < 0.18: [0.36 0.64] Yes
|   |   |   |   ez >= 0.18: [0.89 0.11] No
dz >= 2.11: [0.21 0.79] Yes
```

Figure 1. Determining presence of a pitch accent

To produce a measure of the difference, a C4 type classification tree [5] was grown using questions on the output of the models to partition the ToBI labeled data. 10-fold cross validation, repeatedly holding out 10% of the data and training on the remainder, was performed to estimate the average predictive accuracy from each of the held-out test sets. This method is computationally expensive, but allows all of the data to be used for training, pruning and testing of the tree, and provides an unbiased estimate of prediction performance for unseen data of the same type. Simply growing the tree to completion will result in a much better classification of the training data and shows the degree to which the label set classifies the data, but no inferences could be made about the generalisability of the tree thus grown. Rigorous pruning using held-out data ensures that no such overlearning occurs, and gives a more reliable estimate of prediction by eliminating all parameters that are not productive.

Figure 1 shows a sample tree grown from *wnc* to predict the presence or absence of an accent on a syllable. We can see from the order of the questions that segmental lengthening (dz) and energy on the rhyme (ez) come higher than the average lengthening across the syllable (dzm) and the slope of the energy contour (edm). The numbers in square brackets show the probabilities of each answer (yes or no) determined for each leaf node.

5. RESULTS

In all, 15 tests were performed, using combinations of the *tgt*, *rfc*, and *wnc* models against the 2713 labelled syllables of test data. The combinations were: *tgt* alone,*rfc* alone, *wnc* alone, *tgt+wnc*, *rfc+wnc*, predicting types of i) break indices, ii) phrasal tones, and iii) pitch accents.

Table 4 gives confusion matrices from trees grown on the full data set, without pruning, for the pitch accent predictions from all combinations of the models. Rows in the table show distributions for the predicted labels and columns show those for the original labels. The last column, showing overall predictions, indicates the fit of each model to the data. We can see that the labels are adequate to model the training data well.

```
 none    H+!H     H*      L*      L+H*
(59.8%   2.7%   31.1%     2%      4.2% = 100%)
            wnc alone 93.5% (64.65%)
 0.586   0.001   0.009   0.004   0.003 | 0.605 none
 0.0     0.014   0.0     0.0     0.0   | 0.014 H+!H
 0.010   0.011   0.301   0.005   0.013 | 0.343 H*
 0.0     0.0     0.0     0.009   0.0   | 0.009 L*
 0.001   0.0     0.000   0.0     0.025 | 0.027 L+H*
-----------------------------------------
            rfc alone 67.4% (65.86%)
 0.517   0.016   0.158   0.016   0.015 | 0.723 none
 0.001   0.003   0.001   0.0     0.0   | 0.005 H+!H
 0.078   0.008   0.152   0.004   0.023 | 0.267 H*
 0.0     0.0     0.0     0.0     0.0   | 0.0   L*
 0.000   0.0     0.0     0.0     0.002 | 0.002 L+H*
-----------------------------------------
            tgt alone 73.1% (67.74%)
 0.507   0.019   0.143   0.015   0.017 | 0.703 none
 0.0     0.002   0.0     0.0     0.0   | 0.002 H+!H
 0.088   0.006   0.165   0.001   0.020 | 0.283 H*
 0.000   0.0     0.002   0.003   0.0   | 0.006 L*
 0.0     0.0     0.000   0.0     0.004 | 0.004 L+H*
-----------------------------------------
            rfc + wnc 95.8% (67.09%)
 0.588   0.001   0.005   0.001   0.001 | 0.598 none
 0.0     0.020   0.001   0.0     0.000 | 0.021 H+!H
 0.008   0.005   0.297   0.000   0.003 | 0.315 H*
 0.0     0.0     0.001   0.017   0.0   | 0.018 L*
 0.001   0.0     0.005   0.0     0.037 | 0.045 L+H*
-----------------------------------------
            tgt + wnc 94.2% (67.48%)
 0.588   0.002   0.013   0.005   0.006 | 0.617 none
 0.0     0.022   0.0     0.0     0.0   | 0.022 H+!H
 0.008   0.001   0.297   0.004   0.011 | 0.323 H*
 0.000   0.000   0.0     0.010   0.0   | 0.010 L*
 0.0     0.0     0.0     0.0     0.025 | 0.025 L+H*
-----------------------------------------
 0.598   0.027   0.311   0.020   0.042 = 1.000
 none    H+!H*    H*      L*     L+H*
```

Table 4. Confusion matrices from the full trees (percentages after pruning shown in brackets)

The pattern of predictions in Table 4 shows that without the additional information on segmental lengthening and energy changes, both pitch models tend to miss many of the accents. The first column (unaccented syllables =

59.8% of the data) shows that *rfc* and *tgt* alone predict 72.3% and 70.3% of the syllables as unaccented. When combined with segmental information, this overprediction reduces to 59.8% and 61.7% respectively (end column), which is much closer to the real situation.

On this measure it appears that the combined models (especially *rfc+wnc*) are worth considering further. That is, encoding the slope of the lengthening and energy changes, along with the slope of the fundamental frequency in the locality of a syllable might be an effective way of modelling the higher-level prosodic events in the speech.

	wnc	rfc	tgt	rfc+wnc	tgt+wnc
break index	54.93	43.42	48.39	54.11	55.49
pitch accent	64.65	65.86	67.74	67.09	67.48
phrase tone	86.85	84.92	84.92	87.82	88.45

Table 5. Prediction rates after rigorous pruning.

However, after pruning, the performance of all models drops because of the uneven distribution of tokens in the data. For pitch accent prediction *rfc* is reduced from 644 to 11 leaf nodes, and *tgt* to 15. Table 5 gives average percentage results for all combinations over ten cycles of the test data. The high results for phrasal tones are misleading; the tree learns to predict the most common variants and succeeds by guessing the default (*i.e.*, none) in the majority of cases, and L-L% elsewhere. As expected, duration and energy are better predictors of break indices than is the pitch contour, and the addition of fundamental frequency information does little to improve performance.

Examination of the pruned trees shows that for the two cases of most interest (*rfc+wnc* and *tgt+wnc*) the number of parameters remaining from the f0 models is 2 and 4 respectively. The first and seventh questions for pitch accent prediction in *rfc+wnc* both measure f0 change within the syllable. The third, seventh and eighth in *tgt+wnc* measure target height, and only one question involves the labels. All other questions are as in Figure 1, using lengthening and energy to detect a likely prominence, then putting out only H* decisions.

6. DISCUSSION

Although the intonation models contain enough information to code the contour well, they perhaps specify too fine a level of detail, resulting in too many labels per syllable, with the effect that the trees overlearn the accents not from a generalisable pitch trend, but from unique label combinations, which are not robust against smoothing, leaving only values of absolute pitch change across the syllable as the robust feature.

Similar results can therefore be achieved from simpler inclusion of raw f0 data such as the local average f0 and the amount of f0 change across the syllable. A tree grown using these two parameters (with duration and energy) pruned down to 17 leaf nodes and predicted an equivalent 66.74% for pitch accents (*f0+wnc*).

Repeated testing on 10% subsets of held-out data means that infrequent accent types rarely occur in significant numbers. If we merge accent types (L*, L+H*, H+!H* (total 8.7%) with H* (31.1%)) and measure instead accent detection, then *rfc+wnc* correctly detect 35%, *tgt+wnc* 47% and *f0+wnc* 65% of accented syllables. False detection rates (non-prominent syllables being flagged as accented) are 6%, 13% and 20% respectively.

7. CONCLUSION

The purpose of this study was to determine which representation of the fundamental frequency contour best complements the duration and energy clues for prominence and boundary detection. Two representations were tested; a target model, and a rise/fall model. Both work fully automatically with input representing the raw f0 values, and both encode local contour information by expressing the current status relative to the surrounding values, either as a slope (rise/fall) or as a turning point (top, bottom, rising, falling). Neither was selected, as simple measures of f0 appear adequate.

There may be difficulties inherent in any comparison of different labelling systems that hampered performance of the models, and some of the difference is likely to be due to the fact that a human labeller takes much more context into account than simply the local perturbations in the global f0 when marking a transcription.

I should also point out that neither intonation model tested was designed with such limited use in mind, and no conclusion should be drawn about the individual models themselves from these findings. The Hirst model was used at default settings. It may be advisable under these conditions to readjust the parameters so that the contour specification is much looser, resulting in fewer target points and a more global shape indication for each syllable.

To improve the modelling, it is perhaps necessary to combine the pitch information with that for duration and energy, but in a way that covers a wider context than just the immediate local contour of a syllable. To do this we are currently attempting to develop models at the higher levels of the prosodic word, phrase, and intonation group, and to explore the grammar of these units, following the example of speech recognition, which became more successful at recognising sub-word units when it took a top-down approach, constrained by a grammar, rather than restricting itself solely to the acoustic waveform.

Acknowledgments

The author is grateful to the management of ATR for supporting this research, and to Daniel Hirst and Paul Taylor for making their software available for this study. I trust they will forgive me for putting their models to uses in a way for which they were probably not intended.

References

[1] W. N. Campbell, **Automatic detection of prosodic boundaries in speech** pp 343-354 in Speech Communication 13, 1993.

[2] **ESPS/waves+** , Entropic Research Laboratory, Inc, 600 Pennsylvania Avenue, Washington DC 20003.

[3] D. Hirst, **Un modèle de production de l'intonation**, pp 297-315 in *Travaux de l'Institut de Phonétique 7*, Aix en provence, 1980.

[4] D. Hirst, **Automatic modelling of fundamental frequency using a quadratic spline function**, pp 71-85 in *Travaux de l'Institut de Phonétique 15*, Aix en provence, 1980.

[5] J. Quinlan, **Simplifying decision trees**, pp 239-252 in Gaines & Boose eds *Knowledge Acquisition for Knowledge-based systems*, Academic Press, London, 1988.

[6] K. Silverman *et. al.*, **ToBI: a standard for labeling English prosody**, pp 867-870 in Proc ICSLP Banff, 1992.

[7] P. A. Taylor, **A phonetic model of English intonation** PhD thesis, Edinburgh University, 1992

[8] P. A. Taylor, **The rise/fall/connection model of intonation**, to appear in Speech Communication.

[9] M. Wang & J. Hirschberg, **Automatic classification of intonational phrase boundaries**, pp 175-196 in Computer Speech & Language 6, 1992.

[10] C. W. Wightman & M. Ostendorf, **Automatic labeling of prosodic patterns** to appear in IEEE Transactions on Speech & Audio.