Using Prosodic Structure to Improve Pitch Range Variation in Text to Speech Synthesis

Robert A. J. Clark
Centre for Speech Technology Research,
University of Edinburgh, Scotland

ABSTRACT

The intonation produced by current text-to-speech systems is often either flat or artificial sounding. Pitch range is one of the contributing factors which could be improved by more detailed linguistic knowledge.

In this study, a corpus of read speech is analysed to provide information about prosodic structure and pitch range, which can be used to improve the intonation models for speech synthesis.

The results show how the pitch range variation is most apparent at a *tone group* level of prosodic structure, and how phrase initial and phrase final tone groups have significantly different pitch ranges from tone groups which are phrase medial.

1. INTRODUCTION

One speaker from the Boston Radio News Corpus (f2b) [4] is analysed in this study to determine what part phrasing and pitch range play in the intonation characteristics of this particular speaker.

The prosodic structure of the utterances in the corpus is taken to be: *sentences*, consisting of *phrases*, consisting of *tone groups*. The terms used here for phrase structure units have been chosen to avoid possible incorrect implications that might be associated with some of the more traditional terms. However, the idea of tone group used here is similar to the tone group label used by Ladd [3] (although the implication that it is the same unit should not necessarily be made). The term *phrase* is being used instead of *intonational phrase* [1], again to avoid possible confusion.

One of the reasons that confusion exists between definitions of levels of phrasing is due to the nature of the example utterances which are used as illustrations. Short utterances will hide any complex underlying structure, as one high level of phrasing consists of exactly one phrase unit from the level below it, effectively masking any distinction between these levels. Unfortunately example utterances given in the literature to illustrate phrasing are often of this type.

In this study, the phrase is defined as something equivalent to a length of utterance ending with a extended ToBI[5] break index of 5 [6], and the tone group is defined as a length of utterance ending in a break index of at least 3.

The analysis involves an investigation of statistical differences in pitch range characteristics between tone groups and phrases in different positions within the utterance.

2. MEASUREMENTS

The data, consisting of around 2700 tone groups, was initially automatically labeled for phrases and tone groups. These labels were based on the presence of the ToBI boundary tones and break indices. These labels were then refined by hand, special attention being given to phrases of large numbers of tone groups, as there was usually an obvious phrase boundary which was missing from the automatic labelling.

For each tone group the following measurements were then taken:

$start\_f0$: The f0 value at the onset of voicing at the beginning of the tone group.

end_f0: The f0 value at the cessation of voicing at the end of the tone group.

max_f0: The maximum f0 value within the tone group.

Δ_f0: The difference between the maximum and minimum f0 values within the tone group.

$mean\_f0$: The mean f0 value within the tone group, calculated as $\frac{\sum f0}{n}$ for each pitch tracked f0 point within the tone group.

$std.\_f0.$: The standard deviation of f0 points from the tone group mean.

(min_f0 is omitted as it is a linear combination of max_f0 and Δ_f0.)

The f0 data was extracted from a pitch tracking file created using a super resolution pitch detection algorithm. This was then median filtered with an order seven filter to correct outliers, and f0 values below 100Hz were doubled to correct octave errors. The above variables were then extracted from the data for every tone group in the corpus.

If the start or end of the tone group was unvoiced, the second defined f0 value moving towards the center of the tone group was the value taken (the first point was often found to be a remaining outlier).

Before the data could be statistically analysed, decisions had to be made on how to group the data. Tone group position within the phrase, and phrase position within the sentence can be numbered 1 to n, where n is dependent on the phrase or sentence length respectively. Length here is a measure of the number of lower level units a particular unit is comprised of, for example a phrase of length 3 would be a phrase consisting of three tone groups.

The problem with numbering 1 to n is that phrase final tone groups do not get grouped together, and any property of phrase final tone groups would be lost, as would any property of sentence final phrases. Hence both tone group and phrase position are measured 1 to $n - 1$ and f, where f denotes a unit in final position. Tone groups in phrases of length 1 still pose a problem as they are initial 1 and final f.

The final decision after some exploratory statistical analysis was to group the data into subsets, where each subset consisted of tone groups in phrases of a particular length. This effectively eliminates the above problems.

3. RESULTS

3.1. Overview

The resulting measurements are summarised graphically in figure 1. In the figure, the pitch range of each tone group is represented by a dark grey box enclosed in a light grey box. The light box extends vertically from the minimum to the maximum f0 values found within the tone group. The internal dark grey box extends vertically one standard deviation in each direction from the mean f0, which is represented by a horizontal line. The $start\_f0$, end_f0 and max_f0 are represented as black dots joined by lines. The numbers above the boxes signify the number of cases in that category.

The most obvious feature visible on the graph concerns the first tone group in each phrase. The first tone group in a phrase appears to have a greater pitch range and a higher mean than the other groupings. The f0 mean of the non phrase-initial tone groups is around 165-170Hz, whereas the f0 mean of the phrase-initial tone groups is around 200Hz. This shows that the first tone group may have some special status. The phrase final tone groups appear to be slightly lower than the other categories, but not to the same extent to which the phrase initial tone groups are higher.

It is also quite noticeable that the medial tone groups all appear to be very similar in their characteristics.

The graphs indicate that phrase position in the sentence plays no or little part in determining the pitch range used. The statistical analysis further supports this observation, as described below.

3.2. Statistical Analysis

Two way (phrase position vs. tone group position), multivariate analysis of variance shows that phrase position is not a main effect for any of the independent variables. With this in mind the data is split into subsets to produce statistical results which are more interpretable.

3.2.1. Within subsets analysis. Each subset, as mentioned above, contains only tone groups from phrases of a particular length. Within-subset effects with regard to tone group position can be analysed along with between-subset effects, comparing tone groups with the same position in phrases of different length. This grouping also eliminates the problems with grouping initial and final tone groups together. In the within-subset analysis, tone group position is a main effect, significant at $p < 0.01$ for all the dependent variables (not shown). Table 1 shows repeated contrasts results for this data.

Repeated contrasts show significant differences between adjacent tone groups. The initial analysis suggested that the initial and final tone groups differ from each other, and from medially positioned tone groups, but that all medially positioned tone groups are effectively the same. For this hypothesis to hold we would expect to see a contrast between first and second tone groups and between penultimate and final tone groups.

The contrast results show clearly that the first and final tone groups of a phrase differ from those between them. For all groups, the only repeated contrasts found to be significant are between the first and second tone groups in a phrase, or between the penultimate and the final tone groups.

For phrases of length 2, all of the variables tested showed

		Dependent Variable t-test results					
TGs per phrase	TG type contrast	Start t	End t	Max t	Δ t	μ t	σ t
2	1–2	**5.90**	**8.12**	**16.15**	**11.41**	**16.23**	**11.03**
3	1–2	**3.47**	**2.56**	**9.67**	**7.13**	**10.35**	**5.42**
	2–3	**3.25**	**6.15**	**3.93**	*1.84*	**4.05**	*1.71*
4	1–2	**3.29**	*1.06*	**6.16**	**4.37**	**7.97**	**3.58**
	2–3	*1.05*	*1.13*	*1.09*	*0.85*	*1.92*	*-0.09*
	3–4	*1.46*	**4.04**	**2.13**	*0.41*	**2.09**	*0.98*
5	1–2	**5.58**	**5.76**	**11.82**	**8.44**	**13.67**	**7.46**
	2–3	*1.28*	*1.16*	*-1.03*	*-1.42*	*0.16*	*-1.35*
	3–4	*-1.87*	*0.50*	*-1.29*	*-1.01*	*-0.69*	*-0.67*
	4–5	**-2.26**	*0.14*	**-3.36**	**-2.98**	**-4.67**	**-2.42**
6	1–2	**6.79**	**4.61**	**8.42**	**4.52**	**11.82**	**3.95**
	2–3	*0.47*	*-1.38*	*-1.56*	*-0.79*	*-1.56*	*-1.14*
	3–4	*-1.61*	*-0.25*	*-1.36*	*-0.62*	*-0.52*	*-0.70*
	4–5	*-1.28*	*0.46*	*-0.25*	*-0.32*	*-0.81*	*0.42*
	5–6	**-1.99**	*0.04*	*-1.37*	*-0.71*	**-3.63**	*-0.06*

Table 1: ANOVA repeated contrasts for analysis of phrases of equal number of tone groups. Significant (predominantly at $p < 0.01$) values are shown in bold.

a contrast significant at 1%. For phrases of length 3, the final-penultimate position contrast is lost for Δ_f0 and $std.\_f0$, and the initial–second contrast for end_f0 has dropped to 5%, but the other 9 contrasts remain at 1%. For phrases of length 3 the end_f0 initial–second contrast is no longer significant, and in addition to the loss of the final–penultimate contrast for $std.\_f0$, it has also been lost for $start\_f0$. Phrases of length 5 and 6 have all the initial–second contrasts at 1%, while phrases of length 5 have all but the end_f0 final–penultimate contrast, and phrases of length 6 have only this contrast for $start\_f0$ and $mean\_f0$.

If this data were to show a final lowering effect we would expect the end_f0 contrast to show up as significant between final and penultimate tone groups, but this is not reflected in the data here. Evidence of final lowering can be difficult to capture with automatic pitch tracking with no manual intervention as the speech signal tends to degrade at such points. This especially problematic with natural speech where phrase-final voicing cannot be controlled for in the same way as in examples specifically constructed to measure this phenomenon. This is the most likely reason that we do not find a significant end_f0 contrast. The presence of other effects though, does suggest that phrase final tone groups are special.

We can therefore safely conclude that there are clear pitch range differences between individual tone groups in a phrase, and one such difference specifically manifests itself as a very clear distinction between a tone group which is phrase-initial and those that follow it. There is also a distinction between the pitch range characteristics of a final tone group in the phrase and those that precede it, although this may not always be as clear as the phrase-initial distinction.

3.2.2. Between subsets comparison. We now turn to an across-subsets analysis to see if the pitch range characteristics of the nth tone group in a phrase are affected by the total number of tone

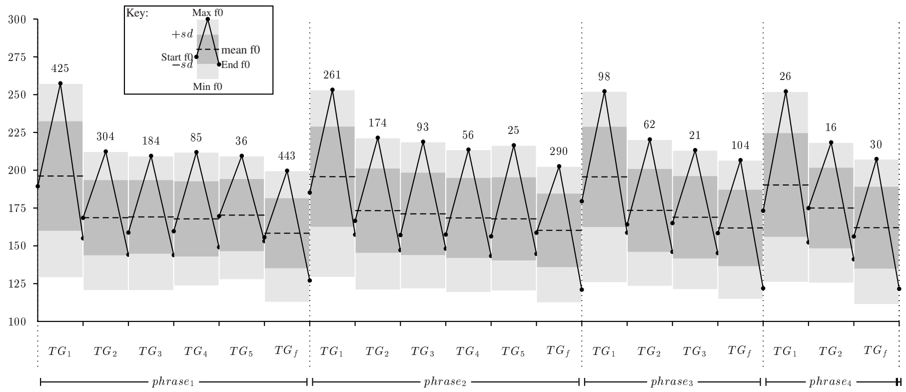

Figure 1: Graphical representation of the pitch range of tone groups and phrases

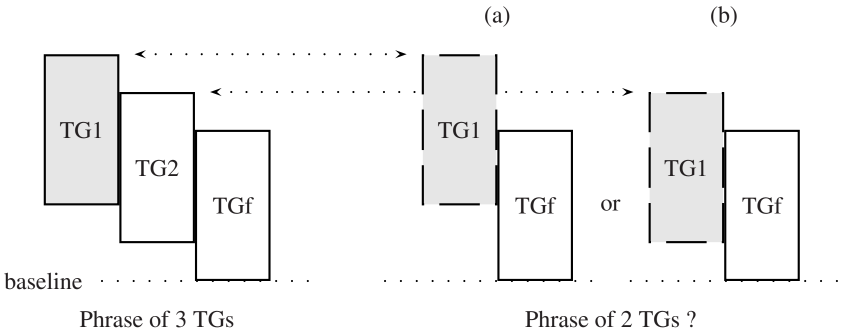

Figure 2: Figure showing two possible alignments for the first tone group in a phrase of 2 TGs compared to the first tone group in a phrase of 3 tone groups.

groups in the phrase.

This analysis concerns the behavior of the overall pitch range structure as the number of tone groups in the phrase is varied. For example (see figure 2): if we compare a phrase consisting of 2 tone groups to one consisting of 3 tone groups, are the pitch range characteristics of the second tone group the same in each case? or are there controlling factors which force them to be different? If we expect that pitch always lowers to the same level at the end of a phrase (there may be multiple phrases to a sentence so we may or may not expect this phenomenon to occur) then we may expect the second tone group which is phrase-final to be lower, or at least finish lower, than the second tone group that is non-final.

The analysis of variance results, for this cross-subset comparison (not shown) reveal in general no interactions between factors, and show $phrase\ length$ to be a clear main effect for most variables. There is also a significant main effect for phrase type for a few variables. There are no significant phrase type contrasts, making an interpretation of this variable difficult. We therefore concentrate on the phrase length main effect.

The repeated measures contrasts, see table 2, show that significant differences only occur between the categories which involve a tone group in phrase-final position. For example: There is only a contrast for the third tone group in a phrase, between phrases of length 3 and 4; here the tone group is in final position in the phrase of length 3, and in non-final position in the phrase of length 4. This seems to be generally true for all of the dependent variables measured.

This reaffirms our hypothesis that a phrase final tone group has special status, and the tone group's finality overrides properties defined by its relative position from the beginning of the phrase. These results also show us that all phrase initial tone groups (excluding those that are also phrase final of course) have the same properties, suggesting that the correct relationship is as shown in figure 2a.

4. RESYNTHESIS MODEL

A crude resynthesis test was used to see if the pitch range distributions that were found are useful for speech synthesis.

The method involved imposing a new mean and standard deviation onto the f0 distribution of each tone group tone of an utterance. A transform was applied to each successive tone group which first normalised the f0, and then mapped the new f0 distribution to

TG		Dependent Variable t-test results					
		Start	End	Max	Δ	μ	σ
pos.	contrast	t	t	t	t	t	t
1	1–2	-1.25	**-3.12**	-1.81	0.21	**-4.48**	-0.60
	2–3	-0.39	**-2.59**	-0.92	0.99	**-2.49**	0.64
	3–4	0.39	-0.74	-0.01	0.31	-1.50	-0.04
	4–5	0.36	-0.38	-0.39	-0.98	0.04	-0.46
	5–6	**-2.28**	-0.53	-0.42	0.61	**-2.41**	0.54
2	2–3	**-2.49**	**-7.36**	**-7.01**	**-4.28**	**-7.17**	**-4.83**
	3–4	0.26	0.89	0.11	0.30	-0.82	0.34
	4–5	-0.46	-0.42	0.59	0.77	-0.67	0.81
	5–6	0.05	1.70	0.32	-0.20	0.52	0.40
3	3–4	0.37	**-7.64**	**-5.55**	**-2.76**	**-5.15**	**-3.30**
	4–5	0.91	0.52	0.95	0.59	-0.95	0.81
	5–6	0.85	0.51	0.31	-0.12	-0.88	0.06
f	1–2	**3.91**	1.06	**9.91**	**9.94**	**6.80**	**8.49**
	2–3	1.20	1.61	-0.93	-1.11	-0.88	-1.61
	3–4	0.74	-0.36	-0.16	-0.48	-0.03	-0.49
	4–5	-0.51	1.19	0.31	0.27	1.27	0.36
	5–6	0.39	-1.54	-0.65	-0.02	-0.41	0.25

Table 2: ANOVA repeated contrasts for cross comparison of tone groups in the same positions in phrases of different lengths. Significant (predominantly at $p < 0.01$) values are shown in bold.

that determined by the new mean and standard deviation.

No analytical comparison of different f0s or perceptual testing has been carried out at this stage, as a framework for the methodology to do this in a meaningful way is still being developed [2]. However, the overall impression from listening to the resynthesised utterances is that the model produces a more varied and 'lively' f0 contour than broader statistical methods produce. This suggests an improvement in pitch range control as long as the resulting liveliness can sound natural.

There is however one noticeable problem with the resynthesised phrase initial tone groups, in that they sound artificially high in pitch when compared to the rest of the phrase. This suggests that the results shown by the tone group analysis do not show enough detail to capture this effect correctly, and that it is not the tone group as a whole that needs to be raised in pitch, but just part of the tone group or particular pitch events within the tone group. Resynthesis using non phrase initial tone group values for the phrase initial tone group removes the artificial 'highness' of the tone group, but loses the initial high which characterises the start of the phrase.

This motivates current ongoing work which is investigating the alignment of pitch events within the tone group. Alignment of events is being considered in terms of their pitch placement with respect to the overall pitch range of the tone group and their time placement with respect to the segmental material in the tone group. Early indications suggest that the raise in pitch is concentrated at the beginning of the tone group, but is not limited to just the first pitch event of the tone group.

5. CONCLUSIONS

We have shown that, of the two levels of phrasing analysed, it is the pitch range of the tone groups that varies with respect to the position of the tone group within the phrase. There is no significant role played by the overall pitch range of the phrase itself.

We have seen that phrase initial tone groups generally have a higher and wider pitch range than non-initial tone groups, and that resynthesis techniques can reflect this phenomenon. A finer level of control in applying this high not to the tone group as a whole, but only to the initial part of it, would improve results further.

We have seen that phrase-medial tone groups generally have the same pitch range characteristics as each other, and can be modeled as such. The fact that their start points are higher than their end points suggests a pitch range reset between tone groups.

Additionally, there are some contrasts between phrase-medial tone groups and phrase-final tone groups, specifically in that they have a lower mean, showing that there is a lowering effect present at the end of the phrase.

Finally, the analysis provides the ground work for further investigation into how particular pitch events within a tone group align, and how they affect the pitch range of that tone group. This investigation is currently under way and the results will hopefully provide a comprehensive model for varying pitch range in this particular type of speech.

ACKNOWLEDGMENTS

Robert A. J. Clark is funded by a PhD studentship from British Telecommunications plc.

REFERENCES

[1] Mary E. Beckman and Janet B. Pierrehumbert. Intonational structure in Japanese and English. In C. Ewen and J. Anderson, editors, *Phonology Yearbook*, volume 3, pages 255–309. CUP, 1986.

[2] Robert. A. J. Clark and Kurt E. Dusterhoff. Objective methods for evaluating synthetic intonation. submitted to EuroSpeech 1999.

[3] D. Robert Ladd. Intonational phrasing: The case for recursive prosodic structure. *Phonology*, 3:311–340, 1986.

[4] M. Ostendorf, P. J. Price, and S. Shattuck-Hufnagel. The Boston University Radio News Corpus. Technical Report ECS-95-001, Boston University, Electrical, Computer and Systems Engineering Department, Boston University, Boston, MA, 1995.

[5] 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*, pages 867–870, 1992.

[6] C.W. Wightman and M. Ostendorf. Automatic recognition of prosodic phrases. In *ICASSP*, volume 1, pages 321–324. IEEE, 1991.