My Voice, Your Prosody: Sharing a speaker specific prosody model across speakers in unit selection TTS

Matthew Aylett, Justin Fackrell, Peter Rutten

Rhetorical Systems Ltd.
matthewa@cogsci.ed.ac.uk

Abstract

Data sparsity is a major problem for data driven prosodic models. Being able to share prosodic data across speakers is a potential solution to this problem. This paper explores this potential solution by addressing two questions: 1) Does a larger less sparse model from a different speaker produce more natural speech than a small sparse model built from the original speaker? 2)Does a different speaker's larger model generate more unit selection errors than a small sparse model built from the original speaker?

A unit selection approach is used to produce a lazy learning model of three English RP speaker's f0 and durational parameters. Speaker 1 (the target speaker) had a much smaller database (approximately one quarter to one fifth the size) of the other two. Speaker 2 was a female speaker with frequent mid phrase rises. Speaker 3 was a male speaker with a similar f0 range to speaker 1 and with a measured prosodic style suitable for news and financial text.

We apply the models created for speaker 2 (an inappropriate model) and speaker 3 (an appropriate model) to speaker 1 and compare the results. Three passages (of three to four sentences in length) from challenging prosodic genres (news report, poetry and personal email) were synthesised using the target speaker and each of the three models. The synthesised utterances were played to 15 native english subjects and rated using a 5 point MOS scale. In addition, 7 experienced speech engineers rated each word for errors on a three point scale: 1. Acceptable, 2. Poor, 3. Unacceptable.

The results suggest that a large model from an appropriate speaker does not sound more natural or produce fewer errors than a smaller model generated from the individual speaker's own data. In addition it shows that an inappropriate model does produce both less natural and more errors in the speech. High variance in both subject and materials analysis suggest both tests are far from ideal and that evaluation techniques for both error rate and naturalness need to improve.

1. Introduction

Prosodic models in unit selection TTS systems have varied from rule based prescriptive models, based on an implicit or explicit knowledge base [1], to data driven models such as: CART decision trees trained from a speakers data [2, 3], lazy learning approaches using tree matching e.g. [4], and unit selection based on a Viterbi search [5]. Prescriptive models have tended to use a neutral declarative prosodic structure which can be dull and wooden to listen to [6], In contrast, statistical models typically suffer from data sparsity problems. In this paper we look at the extent to which we can transfer a data driven prosodic model across speakers to address the data sparsity problem.

1.1. The Model

The prosodic model is based on a lazy learning approach [7], in that duration and f0 values are drawn from examples in the speaker's original data without producing a statistical model. The database consists of an entry for each syllable in the speaker's unit selection database coded for a set of symbolic target features such as syllabic structure and phrase position (similar to the database approach described in [8]). In addition, a set of parametric join features based on the f0 values of the syllabic nuclei to the left and right of the syllable are calculated. Thus two syllables will have a low join cost if they have similar f0 contexts.

This database is used in a *first pass* prosodic unit selection phase (similar to the technique described in [5] where a unit selection approach for prosody generation is applied to a imitation speech database). The results of this prosodic unit selection are used in a *second pass* waveform unit selection phase (See Fig. 1).

As in a traditional unit selection algorithm the same features in the database are calculated for each unit (syllable) in the target utterance. A distance metric calculates a target cost for each syllable and a join cost is calculated during a Viterbi search. The optimal path through the Viterbi lattice is then used to generate duration statistics and an f0 track. These are then used in the, second pass, conventional, unit selection process to produce the final synthesised output.

1.1.1. Duration Targets

The duration for each segment is calculated by taking the duration of the syllable selected from the prosodic database, and distributing it between the target phones while maintaining proportions derived from the segment's mean duration. For example given the second syllable in really /l ii/, if the syllabic duration proposed by the model was 210ms and the mean duration of /l/ was 100ms and /ii/ 200ms we would shrink these segmental durations by 30% to 70ms and 140ms respectively.

1.1.2. f0 Targets

An f0 is constructed by concatenating the interpolated f0 tracks from the selected syllables. f0 targets for each phone are derived from this contour. At present no post smoothing is carried out, instead we depend on the f0 join features in the unit selection process to minimise sharp f0 discontinuities.

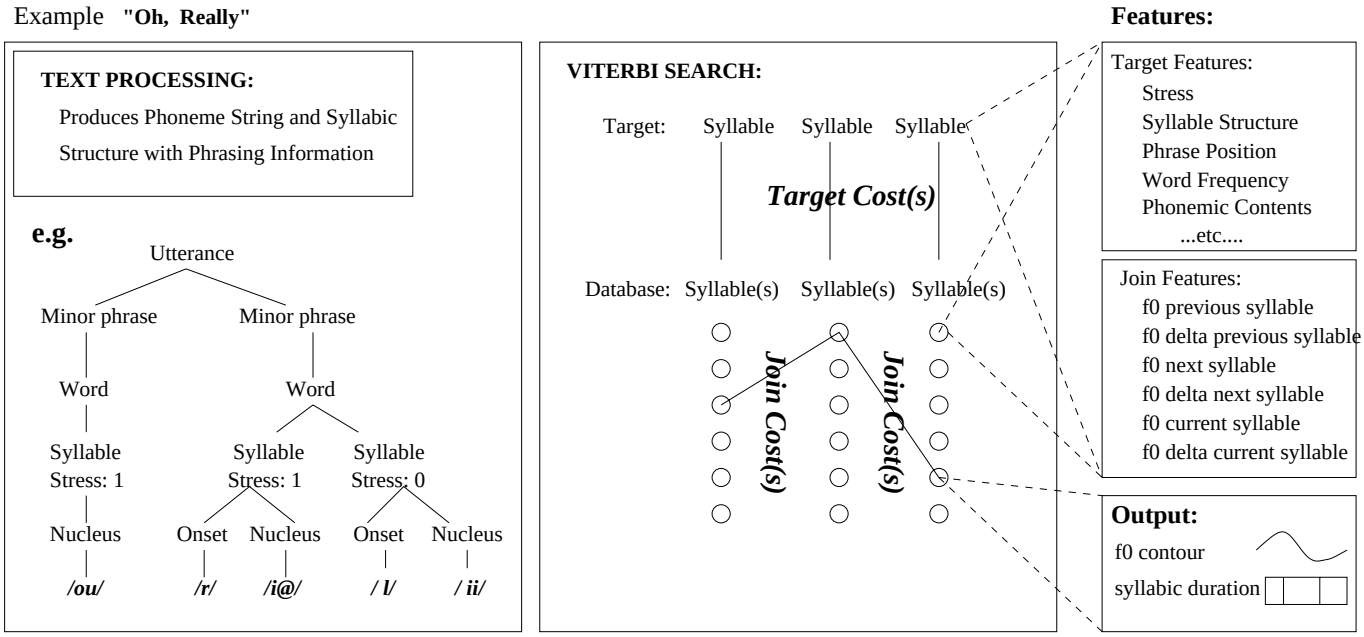

Figure 1: Schematic of unit selection lazy learning prosodic modeling.

1.2. Data Sparsity

A serious problem with this approach is data sparsity. Although this is minimised to a certain extent by reducing the importance of the segmental contents of each syllable ('dad' and 'gag' will be treated almost the same if located in the same prosodic context), the number of distinct prosodic contours in a normal unit selection database is still quite small (hundreds rather than thousands). In order to investigate this problem a prosodic database was generated for a speaker with a small database. In order to avoid DSP artifacts no post unit selection pitch or duration modification was carried out. The target speaker's mean f0 was 134Hz with a standard deviation of 35Hz. The synthesised output of this small sparse database was then compared with the results using much larger prosodic databases from two other speakers: 1) A completely inappropriate female speaker with frequent mid phrase rises, (mean f0 172Hz with a standard deviation of 41Hz) 2) A more appropriate male speaker with a similar f0 range and a measured prosodic style suitable for news and financial text (mean f0 129Hz with a standard deviation of 35Hz). Fig. 2 shows the 3 example target f0 tracks from the utterance 'and still he lay moaning'.

It is difficult to assess the interactions between prosodic models and the results of unit selection. Errors which affect naturalness can be split into prosodic errors, such as insufficient final phrase lengthening, and unit selection errors, such as a spectral mismatch between two concatenated units. However a poor prosodic model will also tend to increase non-prosodic errors. This can be because the prosodic targets match very few long contiguous sections of speech in the waveform database leading to more joins (and potentially more concatenation errors). In addition, if, for example the f0 target is unnaturally high, waveform units, which exhibited pitch doubling because they were acoustically problematic, are more likely to be selected.

Furthermore given an error it is difficult to assess whether it is caused by sparsity in the prosodic model (a stupid target), or sparsity in the waveform database (inability to match a sensible target). There is also no reason why this interaction should have a linear effect on the perception of naturalness. One serious concatenation error might affect naturalness much more than the overall prosody.

To increase confidence in our results, two evaluations were carried out. The first is a standard 5 point MOS naturalness experiment, the second a detailed speech analysis by speech scientists grading errors on a word by word basis.

2. Naturalness Experiment

2.1. Method

Three passages (of three to four sentences in length) from challenging prosodic genres (news report, poetry and personal email) were synthesised using the target speaker and each of the three models (own, appropriate, inappropriate). The synthesised utterances were played to 15 native English subjects and 7 non-native English subjects. Each utterance was rated using a 5 point MOS scale. Where subjects were asked to score - 5: Very natural 4: Natural 3: Neither natural or unnatural 2: Unnatural 1: Very unnatural.

Subjects were allowed to listen to each utterance as many times as they wished before giving a judgment. The order of the utterances was randomised for each presentation.

Agreement between subjects varied. Non native subjects agreed less *(Krippendorff alpha 0.15 [9])* compared to native subjects *(Krippendorff alpha 0.31)*. Because of the low agreement rate non-native data was not used in further analysis.

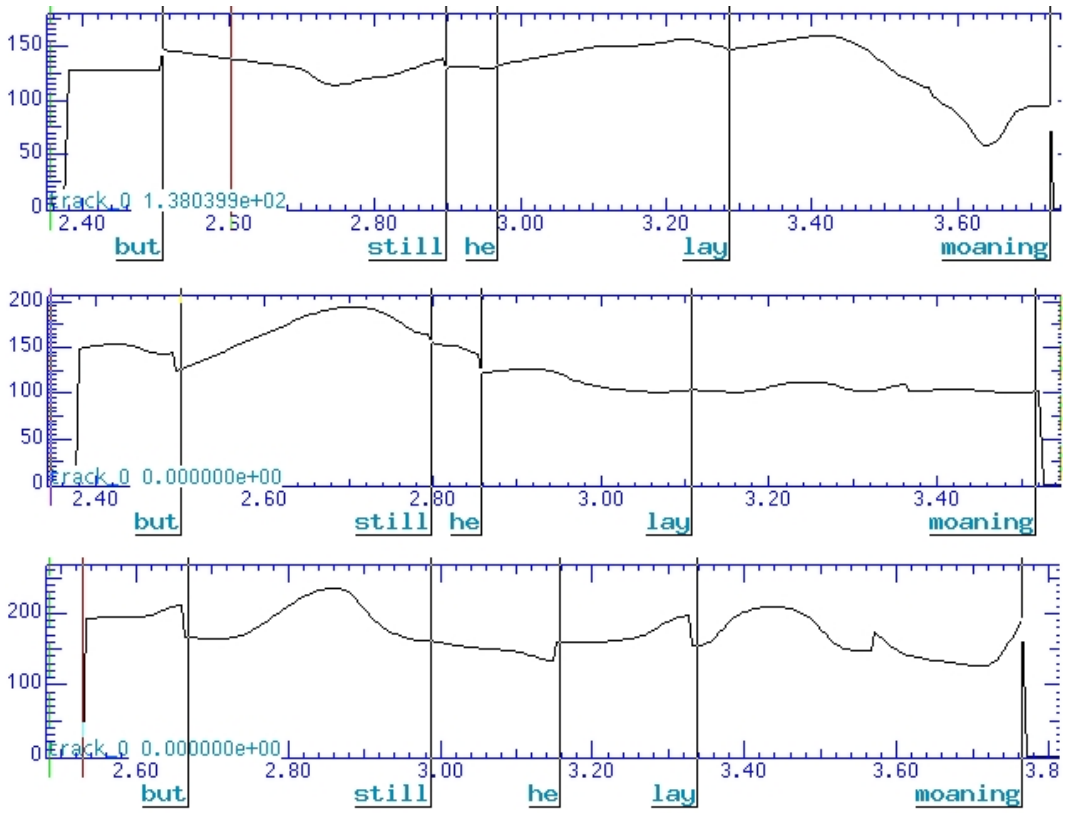

Figure 2: Example f0 tracks generated by the model. Top: An f0 generated by the speaker's own model, Middle: An f0 generated with a large database from a different but appropriate speaker and Bottom: An f0 generated from a large database from a different but *inappropriate* speaker.

2.2. Results

The results (see Fig. 3) were analysed in a by subjects and a by materials ANOVA [10]. Although the model used was significant *(F1(1,28) = 18.94, $p < 0.001$, F2(1,20) = 9.28, $p < 0.005$, MinF'(1,37) = 6.23, $p < 0.05$)* a post hoc test showed that only the inappropriate female model produced significantly worse results *(scheffe, $p < 0.01$)*. Although the mean of the speaker's own model was slightly higher than the model from an appropriate speaker this difference was not significant.

3. Synthesis Error Experiment

3.1. Method

The same passages were used in the synthesis error experiment. 7 experienced speech engineers listened and rated each word on a three point scale: 1. Acceptable, 2. Poor, 3. Unacceptable. They were asked to take into account poor concatenation, mispronunciation, and irregular or inappropriate duration, amplitude and f0 errors. They could listen to the each utterance as many times as they wished. Each word was regarded as a separate data point. The results were analysed in a by subjects and by materials ANOVA. Agreement between coders was equivalent to the naturalness experiment*(Krippendorff alpha 0.34)*.

3.2. Results

The results (see Fig. 4) followed the naturalness results *(F1(1, 12) = 12.81, $p < 0.005$, F2(1, 508) = 6.39, $p < 0.005$, MinF'(1, 27) = 4.26, $p < 0.05$)* and with a significantly more errors on words using the inappropriate model *(scheffe, $p < 0.01$)* and with means where the speaker's own model suggested less errors but again without a significant difference between the appropriate and the individuals speaker's model.

4. Discussion

In general assessing how effective a prosodic model is in a unit selection synthesiser is hard. If the model is given a strong weight you may get more errors because material is not present even though the contour and duration statistics are appropriate. These extra errors will reduce the naturalness. On the other hand if the model makes only a weak contribution to the selection algorithm you can argue that the contour is acting as a guideline only and any differences between models are irrelevant. In this case, the resulting prosody may be poor despite the model, and also produce poor naturalness scores.

A major problem in furthering this work is the time and resources required for assessment. In many cases a bad prosodic model only becomes apparent when the subject is exposed regularly to the synthesis. It is far from clear how this could be assessed efficiently enough to allow sufficient experiments to guide development.

To further complicate matters the assessment can be very subjective. We have carried out a comparative listening test

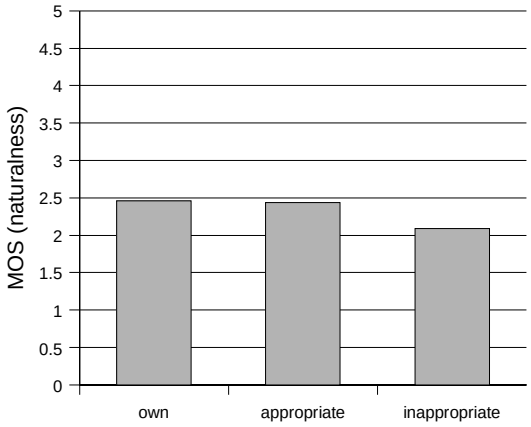

Figure 3: Results for naturalness (by subjects)

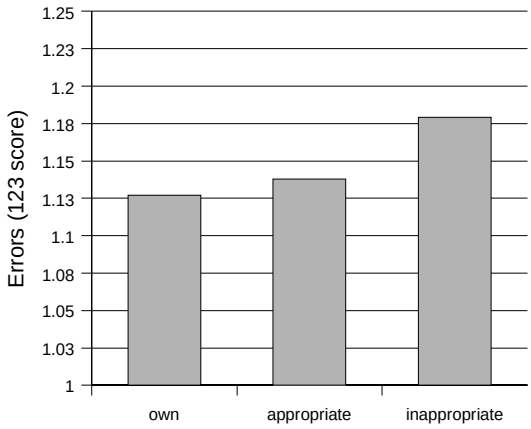

Figure 4: Results for synthesis errors (by subjects)

where one subject chose only the stimuli for which he heard a clear prosodic improvement for one category and avoiding stimuli where preference was based on concatenation or waveform selection errors. When these stimuli where listened to by 3 further subjects no significant preferences were shown. The very low Krippendorff alphas for all the evaluations (the number can be interpreted as approximately 30% agreement) also demonstrate that reliability is a serious problem in such evaluations.

However a poor prosodic model does certainly produce poor synthesis. One reason for this is probably dues to the model requiring waveform units which are not present in the database contiguously. This in turn produces more concatenation and possibly more concatenation errors. In the above experiment the average length of the contiguous unit was highest using the speaker's own prosodic database, slightly lower for the appropriate prosodic database and shortest for the inappropriate database. However, there is not clear relationship between the average length of contiguous units in a synthesised utterance and its perceived quality, especially if you use a good algorithm to join and select units.

The prosodic unit selection approach is an interesting one for three primary reasons:

1. You would hope it would tend to increase the average length of contiguous units (which we might hope is a good thing)
2. We can use automatic weight tuning algorithms with more confidence than in waveform unit selection. This is because we can produce a clear error statistic using copy synthesis. For example given that set of weights how close was the f0 contour to utterance outside the database by the same speaker.
3. The prosodic contours and duration targets produced will vary considerably between utterances and may produce a more natural sounding variation in the speaker's prosody compared to that produced by a prescriptive model.
4. The model can potentially be shared across speakers.

5. Conclusion

The results suggest that a large model from an appropriate speaker does not sound more natural or produce fewer errors than a smaller model generated from the individual speaker's data. In addition it shows that an inappropriate model does produce both less natural and more errors in the speech.

The naturalness results followed the error results very closely, suggesting that basic synthesis errors, rather than style of prosodic expression, are the main contributory factor in any naturalness assessment.

6. References

[1] M. Anderson, J. Pierrehumbert, and M. Liberman, "Synthesis by rule of english intonation patterns," in *ICASSP*, 1984, pp. 281–284.

[2] J. Fackrell, H. Vereecken, C. Grover, J. Martens, and B. V. Coile, "Corpus-based development of prosodic models across six languages," in *Improvements in Speech Synthesis*, E. Keller, G. Bailey, A. Monaghan, J. Terken, and M. Huckvale, Eds. Wiley, 2002.

[3] K. Dusterhoff, A. Black, and P. Taylor, "Using decision trees within the tilt intonation model to predict f0 contours," in *Eurospeech*, 1999, pp. 1627–1630.

[4] L. Blin and L. Miclet, "Generating synthetic speech prosody with lazy learning in tree structures," in *CoNLL-2000 and LLL-2000*, 2000, pp. 87–90.

[5] J. Meron, "Prosodic unit selection using an imitation speech database," in *4th ISCA Workshop on Speech Synthesis*, 2001, pp. 53–57.

[6] D. Jurafsky and J. Martin, *Speech and Language Processing*. New Jersey: Prentice Hall, 2000.

[7] C. Atkeson, A. Moore, and S. Schaal, "Locally weighted learning," *Artificial Intelligence Review*, vol. 11, pp. 11–73, 1995.

[8] F. Malfrère, T. Dutoit, and P. Mertens, "Fully automatic prosody generator for text-to-speech," in *ESCA/COCOSDA Workshop on Speech Synthesis*, 1998.

[9] K. Krippendorff, *Content Analaysis*. Newbury Park: Sage Publications, 1980.

[10] D. Howell, *Statistical Methods for Psychology (5th Edition)*. Duxbury Press, 1997.