

MERGING DATA DRIVEN AND RULE BASED PROSODIC MODELS FOR UNIT SELECTION TTS

Matthew Aylett

Rhetorical Systems Ltd.

ABSTRACT

Data driven models suffer from data sparsity and can be difficult to generalise. Rule based models suffer from being over prescriptive and insensitive to the contents of the unit selection database. To further complicate matters the space of acceptable prosody for any one utterance is large. However in some cases prosodic patterns for a particular speaker can be very homogeneous, for example the prosodic pattern used to read out a zip code. In this paper we describe a method for exploring and analysing the prosodic space within a limited domain, and a method for merging a simple rule based prosodic model with a set of data driven mini prosodic models. A listening test was carried out on the synthesis of zip codes with and without the mini models with promising results. The approach could be applied effectively to domains varying from numerical amounts to personal names.

1. INTRODUCTION

Prosodic models in 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. Two possible solutions to this problem are: 1) To merge data across speakers to reduce data sparsity [7, 8], 2) To combine a prescriptive model with a data driven model. This paper focuses on the second solution with special regard to limited domain synthesis.

Commercial open domain speech synthesis is often applied to particular limited domain tasks (For example noting change of address, reading out banking details). Often these tasks require special care with regards to the intelligibility and quality of synthesis because the speech being synthesised is informationally dense.

Within limited domain synthesis data driven prosodic modeling can be especially successful. For example the

prosody a speaker uses to read a zip code will be more homogeneous than that used in free text. However it is not possible to limit such synthesisers to this limited domain and they must also perform across free text. With this in mind we present:

1. A simple method for analysing the duration and f0 characteristics of a speakers database based on semantic abstraction of word type (e.g. Digits, Letters, Other Numbers), and focusing on the characteristics of the stressed syllabic nucleus of each word.
2. An example of a simple data driven model generated from this analysis.
3. An example of an algorithm to merge this model with a simplistic prescriptive model.
4. The results of applying this algorithm for the target f0 contour of a synthesised utterance.
5. The results of a listening test with this method applied to the synthesis of UK zip codes.

Finally we will discuss this approach might be extended, its limitations and propose further work.

2. ANALYSING SPEAKER SPECIFIC LIMITED DOMAIN PROSODY

Prosodic analysis of speech varies from phonological approaches where speech is hand coded for prosodic categories (e.g [9]) to a parameter based approach where actual f0 and duration statistics are collected from digitised waveforms using autosegmentation and f0 extraction algorithms. The problem with the first method is that hand coding is time consuming and potentially subjective (e.g About a 70% agreement on the presence/absence of a pitch accent between experienced coders [10]).

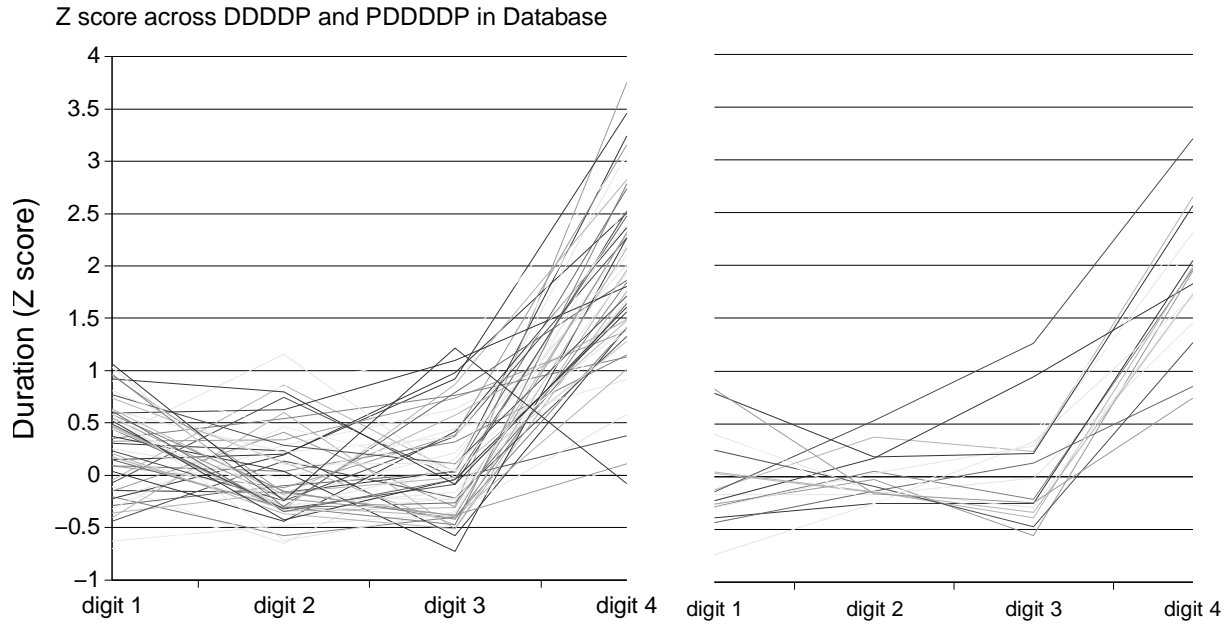


Figure 1

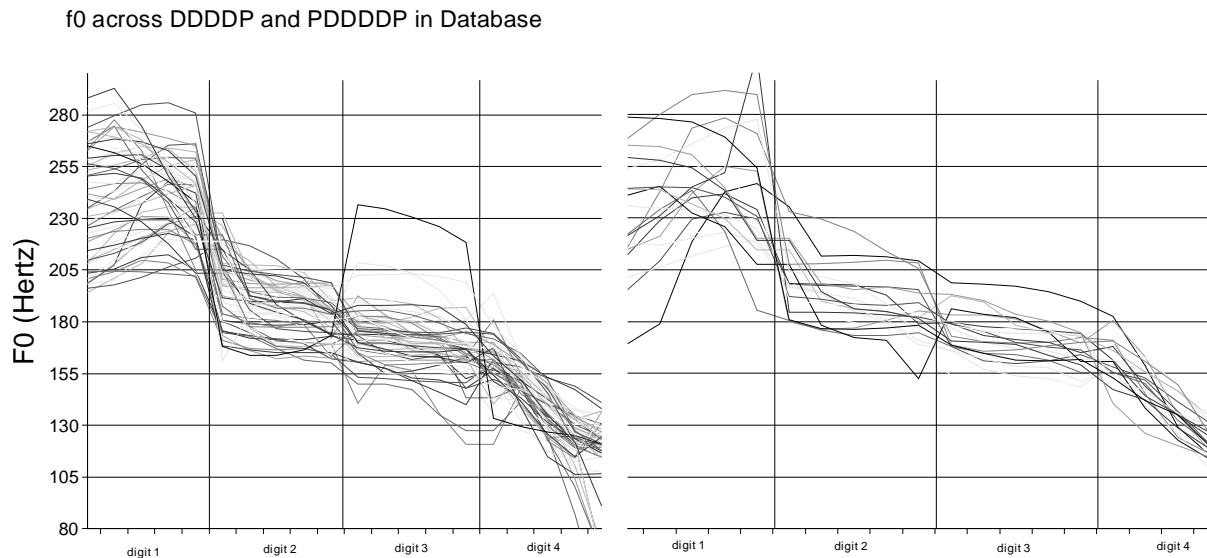


Figure 2

However a parameterised approach suffers from the fact that much of the variation of f0 and duration within a speaker can be unmarked, that is, have little impact on the perceived prosodic structure. This raises the significant problem of deciding what variation to statistically take account of and what to ignore. Given that the objective was to quickly produce a limited domain model for a specific speaker we decided the parameterised approach was more appropriate. In order to reduce the unmarked variation we examined only the nucleus of the primary lexically stressed syllable of each word. This resulted in two sets of measurements:

Duration: The z score of the nucleus's duration measured against the mean and standard deviation of the phonemes duration over the speakers database.

f0: Five f0 values taken from 0%, 25%, 50%, 75% and 100% though the nucleus.

To further reduce the scope of the model we limited the measurements to sequences of words which fell into simple semantic categories:

Letters: (L) Single Letters such as those in the name "I B M".

Digits: (D) Single digits such as 1,2,3 but not natural numbers greater than 10 such as eleven, twenty, hundred.

Other numbers (N, n): Such as hundred, twenty, eleven, point, minus and a special category for the "and" used to speak numbers such as "thirteen thousand **and** eleven".

In addition we marked the presence or absence of a pause before or after the words in the database with the symbol 'P'. All other words were ignored (marked ".") in the example below). The result was to take sentences from the database and produce a symbolic string which represented the limited domain we were interested in (in this case numbers and acronyms), we will refer to these as NDL (number digit letter) sequences throughout this paper.

Examples:

Input: "IBM later reported a rise of 3.8 cents, in its share price."

Normalized: "I B M later reported a rise of three point eight cents *pause* in its share price"

NDL sequences: PLLL.....DND.P...P

Input: "You owe me, \$234.34"

Normalized: "You owe me *pause* two hundred and thirty four dollars and thirty four cents"

NDL sequences: P...PDNnND..ND.P

Input: "Arizona, 12345-1234"

Normalized: "Arizona *pause* one two three four five *pause* dash one two three four"

NDL sequences: P.PDDDDDP.DDDDP

Sequences of 3 or more NDL items were then regarded as valid prosodic mini domains (e.g the "PLLL" in the first example is a phrase initial sequence of three letters.) The frequency of these sequences occurring in the database were measured and then we plotted the Duration and F0 data to examine how homogeneous they were.

Figure 1 shows the zscore duration for the sequences DDDDP and PDDDDP (1 value per lexically stressed nucleus). As we can see the duration change is reasonably stable across examples despite very different phonetic contents.

Figure 2 shows the f0 values taken from the nucleus for the same sequences (5 values per lexically stressed nucleus). As with duration the f0 contour is reasonably stable and minimally affected by whether there was a preceding pause or not¹.

¹The data item with the pronounced stress on the 3rd digit is a repeated item '3131'

Of course other sequences may not be as homogeneous. The DDDD sequence was almost exclusively connected to zip codes in our data. General numbers produce a much more varied set of sequences and potentially much more prosodic variation. We will return to this issue in our discussion.

3. BUILDING A SPEAKER SPECIFIC LIMITED DOMAIN PROSODIC MODEL

Given a homogeneous analysis of an NDL sequence a simple statistical model was built by calculating the mean of each item in the sequence across the same sequences and using this as a target for synthesis. For example if we have 30 PLLL sequences the average z duration of the nucleus of the lexically stressed syllable of the first letter is averaged across all the 30 examples, then the average z duration of the second letter, etc. The same process is carried out for the 5 f0 values in each item. This is the simplest model we could envisage and the one these experiments are based upon.

4. MERGING AN NDL MODEL AND A PRESCRIPTIVE RULE BASED MODEL

The analysis technique has left a significant gap in model when it comes to applying the values in normal synthesis:

1. How do we calculate the targets for material which is not in an NDL sequences?
2. How do we calculate a target for segments which are not the nucleus of the lexically stressed syllable?

The answer to question one is to use a prescriptive model (or your favorite statistical model) to generate values. For f0 we cut the contour produced, splice in our NDL model and then continue the model afterwards. We allow pauses to break the contour and if no pause is present we connect the contours together altering the absolute values to take into account downdrift. For duration we generate completely separate values using the alternative general model.

For the *gaps* in the model we take two approaches:

For f0 we connect the nucleus values to all other values with a straight line interpolated from one to the next.

For duration we compare the **difference** between the NDL specified z duration and the general models z duration and linearly adjust the general model across the whole word until the z duration for the stressed nucleus agree.

For example if we have the word "seven" and the general model predicts a z score of 1.0 for the /e/ in the first syllable and the NDL model predicts a value of 1.2 we increase the z durations for all other segments by 20%.

Comparison of f0 models with database

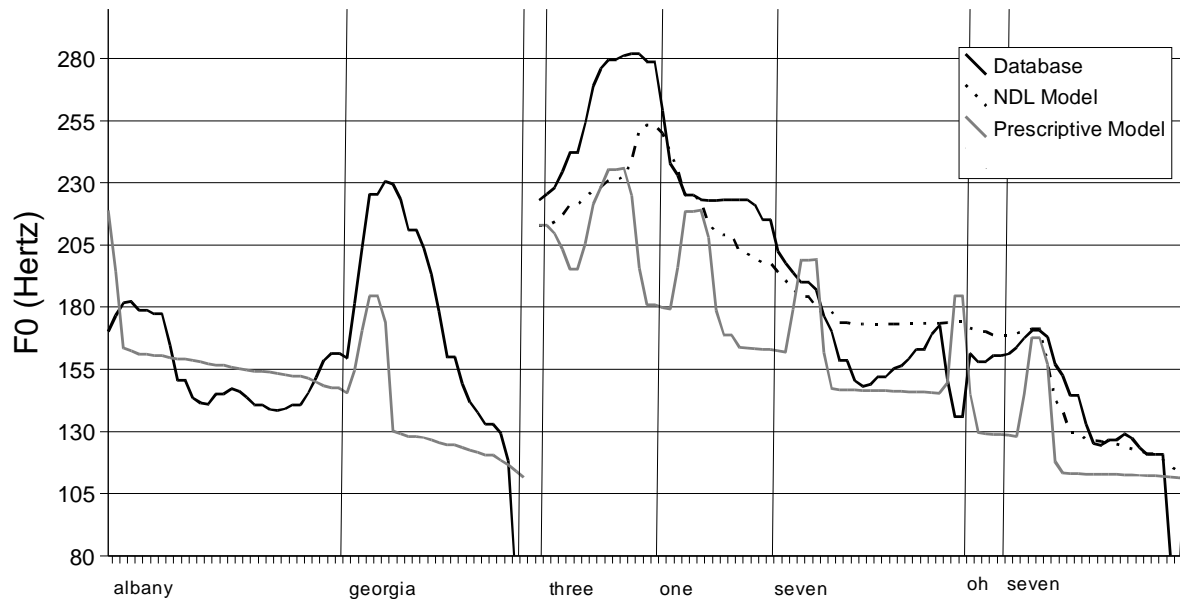


Figure 3

5. EXAMPLE OF A MODIFIED F0 CONTOUR

Figure 3 shows the f0 contour from a database utterance compared to a straw man prescriptive model and with the NDL model for a US zip code.

The straw man model is as follows. For each stressed syllable in a non function word locate a rise fall picture accent. Over each phrase cause a downdrift from 220Hz to 120Hz.

The difference highlights the utility of the mini prosodic model for US zip codes. It closely approximates the f0 in the database. Of course the prescriptive model could also be made arbitrarily more complex in order to match the speakers prosodic patterns. Even so the US zip code mini prosody would remain a good fit.

6. NDL LISTENING TEST

An A/B comparison listening test was carried out on a British RP voice speaking sentences of the form "The postcode is EH1 4ET" or "The postcode is CG12 8LF". which mapped onto to two NDL sequences PLLDPDLLP and PLLNPDLLP. 32 sentences were synthesised with and without the NDL model and using the straw man prescriptive model.

4 subjects listened to both examples of each sentence up to a maximum of three times and carried out a blind comparison test selecting a strong preference, weak preference or no preference for each.

The results were as follows:

Strong Preference NDL	17
Weak Preference NDL	41
no preference	70
Weak Preference Baseline	19
Strong Preference Baseline	2

(Significant $P < 0.005$ Wilcoxon signed rank test)

It is important to note that the improvement was not caused by selecting more material from ZIP codes in the database. In the baseline, 207 of the 596 units used to synthesise the ZIP code portion of the synthesis, were from original zip codes in the database. In the NDL version this remained almost identical (209). This suggests that improvement was due to selecting more appropriate units in terms of f0 and duration.

7. DISCUSSION

As discussed in [7] listening test results for prosodic model changes within unit selection must be treated with care. Other synthesis errors have a heavy impact on results and the extent the prosodic target is taken into account by unit selection can vary dramatically between different systems.

However in this case, because of the limited nature of the sentences being synthesised these results can be more

strongly attributed to prosodic model differences as the general quality of synthesis was very good with few concatenation errors.

A more complex question is to what extent you can generalise these results. Two weaknesses need to be addressed:

- The domain tested was too limited.

Certainly you may not get such good results with less homogeneous mini domains. In addition it is far from clear how you might generalise mini domains further. However given a set of semantic abstractions such as NDL it is possible to maximise homogeneity and generality automatically. Thus we believe the framework outlined here is more tractable than requiring design of prosodic targets for ZIP codes by an expert intonation phonologist. Getting ZIP codes to sound right may seem trivial in terms of open domain speech synthesis but it can be quite important if you want to synthesise thousands of addresses every day to customers over phone lines.

- The straw man model was so poor any change would have been significant.

If the prescriptive model is very good we accept the use of a data driven mini model becomes unnecessary. However, producing a prescriptive model which is appropriate for all limited domains and general synthesis is a non-trivial task. The approach outlined here allows us to plug gaps or weaknesses in any prescriptive model and feel confident that our zip code will be correctly modeled. As such it must be regarded as a straight forward engineering solution to a complex problem, not a reason to avoid improving prescriptive models.

8. CONCLUSION

We have outlined a framework for analysing and modeling mini prosodic domains using number, digits and letters. The approach of merging these mini models with a prescriptive model is potentially powerful. A small listening test focusing on ZIP codes suggests the approach is a practical solution to a tricky prosodic modeling problem and could lead to significant improvement in the prosodic naturalness of unit selection synthesis.

9. 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. Van 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.E. Dusterhoff, A.W. 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.H. Martin, *Speech and Language Processing*, Prentice Hall, New Jersey, 2000.
- [7] M.P. Aylett, J. Fackrell, and P. Rutten, "My voice your prosody: Sharing a speaker specific prosody model across speakers in unitselection tts," in *Eurospeech*, 2003.
- [8] B. Gillet and S. King, "Transforming f0 contours," in *Eurospeech*, 2003.
- [9] Mary E. Beckman and Gayle M. Ayers, *Guideline for ToBI Labelling.*, 1.5 edition, 1993.
- [10] C. Mayo, M. Aylett, and D. Ladd, "Prosodic transcription of Glasgow English: An evaluation study of glatobi," in *Proceedings of an ESCA Workshop: Intonation: Theory, Models and Applications.*, A. Botinis, G. Kouroupetroglou, and G. Carayannis, Eds. October 1997, pp. 231–234, ESCA and The University of Athens.