EUROSPEECH '97
5th European Conference on Speech
Communication and Technology
Rhodes, Greece, September 22-25, 1997

DATA DRIVEN FORMANT SYNTHESIS

Jesper Högberg
Department of Speech, Music and Hearing, KTH,
S-10044 Stockholm, Sweden
Tel. +46 8 790 78 94, FAX: +46 8 790 78 54, E-mail: Jesper.Hogberg@speech.kth.se

ABSTRACT

In this study we introduce combined data driven and rule based methods to synthesise speech. The aim is to improve on the coarticulatory modelling by adapting the KTH TTS system to data from one speaker. Regression trees are trained on a manually corrected speech database to provide predictions for vowel formant frequencies. At runtime, the TTS system produces formant frequency trajectories that are derived from weighted contributions from both the rules and the regression trees. The weighting strategy allows flexible adjustment of the synthesis parameters and thus of the quality of the output speech. An informal perceptual test was conducted to compare the performance of the hybrid approach to that of the traditional rule based system. A great majority of the test subjects judged the speech output of the hybrid system to be more natural than the competing rule derived speech. The speech produced by the hybrid system was also generally preferred.

10.21437/Eurospeech.1997-210

1. INTRODUCTION

Data driven speech synthesis has gained much attention in recent years. Concatenative techniques, like for instance the PSOLA-technique, [1] are, by now, well established. The same techniques that have been developed in large vocabulary speech recognition have been successfully applied to derive speech synthesis units from a corpus of continuous speech [2]. The advent of concatenative speech synthesis techniques has meant a breakthrough in segmental naturalness. However, spectral mismatches at joints remain a problem and the limited flexibility inherent in non-parametric approaches are problems that are hard to solve. Lately, the interest for knowledge based formant synthesis has declined. Speech produced by formant synthesis, though highly intelligible, often suffers from lack of naturalness.

This investigation introduces data driven formant synthesis, which combine some of the advantages of both data driven and rule based speech synthesis. Our goal is to build a system that can be adapted to an arbitrary speaker using a speech database. In our approach, a data driven method to predict synthesis parameters has been implemented in the rule based TTS system developed at KTH [3]. Regression trees are trained on a manually segmented speech database from one male speaker. The database also includes formant frequencies which are used in the training of the regression trees. The system makes use of weighted contributions from rules and tree predictions to produce formant frequency trajectories. In this way we try to improve the coarticulation modelling in the TTS system. An informal perceptual test was set up to evaluate the new hybrid system.

An overview of the TTS system is given in section 2 followed by a description of the speech data used for training of the system. Descriptions of the experimental procedure and the results of an informal perceptual evaluation are given in sections 4 and 5. The last section contains some conclusions.

2. SYSTEM OVERVIEW

The KTH TTS system utilises definitions of language specific phonemes that are defined in terms of phonetic features and default synthesis parameter values. Phonetic rules modify the synthesis parameters based on the default values. In particular, formant frequency trajectories are determined by interpolating between defined target values. Whether a target is reached or not depends on the values of transition-speed parameters and the segment duration. Finally, the parameter values are passed to a formant synthesiser to produce speech. In the data driven mode, the rule based method of setting vowel formant frequencies is complemented with predictions from regression trees. The regression trees are derived from training data prior to synthesis. Figure 1 shows a schematic overview of the system.

2.1. Regression trees

Classification and regression trees [4] are widely used in speech recognition to cluster acoustic data into homogeneous sets, e.g. [5]. Recently, the same technique has also been used to select suitable speech synthesis units [2]. In this investigation we derive regression trees that are used to predict synthesis parameter values for vowels.

The training data is partitioned into two subsets by asking questions about the target phone and the phonetic context. The data responding positively to the question are assigned to the first subset while those responding negatively are assigned to the second. A question can, for instance, be *"Is the following phone a velar?"* The question that minimises a cost function is used to define a partition of the data into two subnodes. The tree is grown by recursively partitioning the data into new subnodes until a stop criterion is met.

Data from all the phones to be analysed can be pooled initially and then forced partitions can be made on the

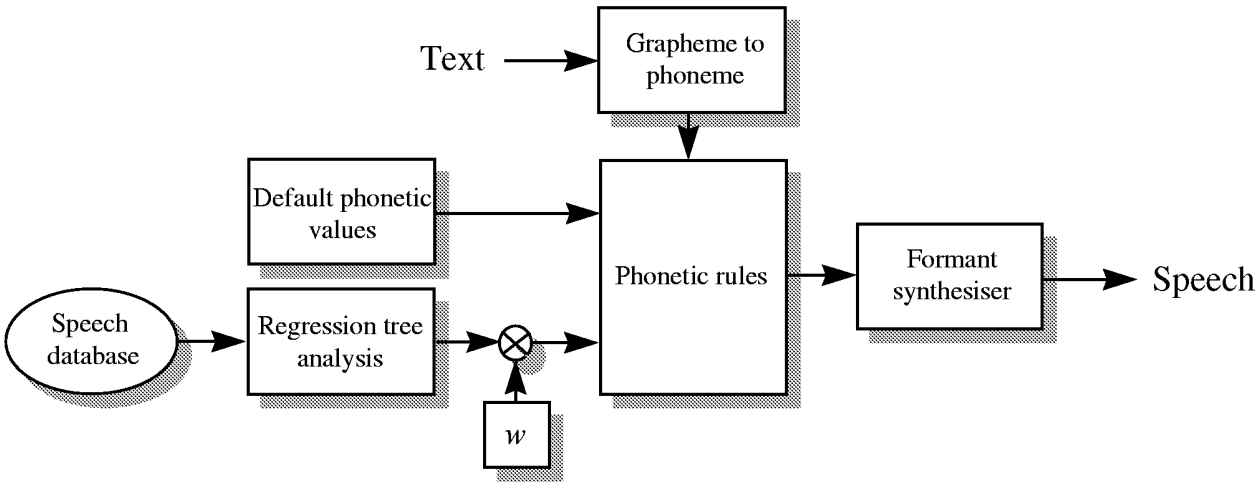

Figure 1. Block diagram of the modified KTH TTS system.

target phones to obtain tree branches that contain data of only one phone. Large trees that accommodate predictions for all phones can thus be derived. Omitting forced partitions with regard to the target phone provide acoustic models that can be shared by several different phones in a specified context [9]. Consider, for instance, the case of diphthongs. These phones can share the same acoustic model for the centralised offset if they prove to be acoustically similar.

2.2. Synthesis procedure

When the system is invoked in the data driven mode the regression trees are loaded. When a vowel phoneme is found in the transcription string it is passed to the regression tree prediction module with information about neighbouring phones, segment duration, etc. The tree is traversed following the path to a leaf node by responding to the questions in the nodes about the context etc. In the leaf node phonetic predictions, e.g. formant frequency values, can be determined. The combined rule and data driven approach is quite flexible. Synthesis parameters can either be set according to rules and, thus, be used as predictor variables for other parameters, or they can be predicted from data. Segment duration, for instance, can be set by the rules or by predictions from the training corpus.

The rule based system generates synthetic speech corresponding to elaborate articulations, approaching the hyper extreme in the hyper-hypo dimension [10]. On the other hand, a large part of the segments in the speech database are reduced and could accordingly be described as hypo-like realisations. In order to model these variants continuously, formant frequencies can be set to values derived from weighted contributions from both the rules and the regression trees. The data derived formant frequency predictions are weighted with a value provided by a parameter *w*. Consequently, the rule generated formant frequencies are given the weight 1-*w*. One advantage of this procedure is that transitions between rule and data derived parameter values can be smoothed. This is especially useful in the current study where only the formant frequencies of the vowels are predicted from data. Acceptable coherence between vowel and consonants can be achieved by adjusting *w* to an appropriate value.

Running the system in the data driven mode also means that a different formant frequency interpolation scheme is used. The formant transition speed parameters are adjusted to ensure that the predicted values are actually reached at the relative time in the segment where they were measured. However, the maximum transition time is set to 40 ms. Therefore, segments with long duration will have a stationary part. This means that the data driven mode, with *w* set to zero, does not produce formant transitions identical to the ones of the rule based mode since the respective interpolation schemes differ.

3. SPEECH MATERIAL

The speech data used to derive regression trees pertains to one male subject reading eleven short stories: corpus one, and newspaper text: corpus two. Both corpora have been manually segmented and labelled prior to this investigation. Corpus one has been used in several other studies, e.g [6] and [7]. The corpora contain more than 30 minutes of speech. In all, the speech data comprise some 12.000 vowels. The first four formant frequencies were estimated automatically, using the ESPS/waves+ formant tracking facilities [8] and, subsequently, hand corrected.

4. EXPERIMENTAL PROCEDURE

An informal listening test was designed to assess the performance of the combined rule based and data driven system compared to that of the traditional system. Section 4.1 describes the derivation of the regression

trees used in the experiment and section 4.2 describes the perceptual test conditions.

4.1 Regression tree analysis

The entire speech data set, consisting of both corpus one and corpus two, was used for training. One regression tree was derived to predict the first three formant frequencies of vowel onsets (the first frame in the segment). Two more trees were constructed to account for formant frequency predictions for the vowel centres and offsets. The trees accommodated data from all vowel phones but with separate branches for each phone. The predictor variables were the immediate left and right contexts. The question set contained 147 questions on single phones and broad phone classes. In each node the objective function R was minimised

$$R = R_l + R_r$$

$$R_l = \frac{N_l}{N_l + N_r} \sum_{i=1}^{N_l} \sum_{j=1}^{3} \left(F_{i,j} - \bar{F}_j \right)^2$$

$$R_r = \frac{N_r}{N_l + N_r} \sum_{i=1}^{N_r} \sum_{j=1}^{3} \left(F_{i,j} - \bar{F}_j \right)^2$$

where N_l and N_r are the number of samples in the left and right node respectively. $F_{i,j}$ are mel-scaled formant frequency vectors and $\bar{F}_j$ is the mean mel frequency of the jth formant in that node. Thus, R is the sum of the squared mel formant frequency deviations from the means in the left and right node weighted by their relative node sizes. Potential partitions yielding fewer than three samples in either of the two subnodes were disregarded. Partitioning was stopped when a node contained less than nine samples or when the context was fully specified and no further partitions were possible. There were 1360, 1381 and 1383 leaf nodes in the vowel regression trees for the onset, centre and the offset of the vowels respectively. Each leaf node contained about nine samples on average. The formant frequency vector at the minimum squared Euclidean distance from the centroid of the leaf node was used for prediction.

4.2 Perceptual evaluation

A few sentences from corpus one was synthesised in three different versions. The first version, V1, was synthesised using the traditional rule based system. The second version, V2, was synthesised in the data driven mode using the regression trees described in the previous section. The parameter w was set to 0.6. (Hence, the tree predicted formant frequencies were given the weight 0.6 and, accordingly, the rule derived formant frequencies were given the weight 0.4). The same five sentences were used for V1 and V2. A third, and longer version, V3, was synthesised at an intermediately reduced level with w set to 0.2.

All versions were synthesised using the same speaking rate in order to make them different with respect to formant frequencies only. This speaking rate was somewhat higher than the normal rate used in the TTS system. All segment durations were scaled with a factor 0.9 to achieve the increase in speaking rate. This choice was made as a compromise between the normal speaking rate produced by the system and the natural speaking rate used by the male speaker for these utterances. The specifics of the different versions are summarised in Table 1.

VERSION	SYNTH MODE	# UTTERANCES
V1	Rules	5
V2	Data, $w = 0.6$	5
V3	Data, $w = 0.2$	15

Table 1. Summary of the synthesised versions used in the perceptual test.

Opinions regarding the quality of V1 compared to that of V2 were collected from fifteen people. Two categories of listeners participated in the perceptual evaluation, one category consisted of ten "inexperienced listeners" that were not used to synthetic speech. The remaining five subjects were "experienced listeners" used to the speech produced by the rule based system. The experienced listeners were not, however, aware of the aim of this study.

All subjects listened to V3 in a training session. The purpose of this session was primarily to acquaint the inexperienced listeners with synthetic speech. The intermediate reduced quality of V3 was applied to avoid biasing of the subjects in favour of V1 or V2. Following the training session, the subjects listened to V1 followed by V2 at least three times. The subjects were then asked to answer two questions:

- *Which version, V1 or V2, sounds most natural?*
- *Which version, V1 or V2, do you prefer?*

The subjects were also encouraged to supply additional comments. No subject knew in advance what the difference between V1 and V2 consisted in.

5. RESULTS AND DISCUSION

The overall voting scores for V1 and V2 with respect to naturalness and preference are given in Table 2. All subjects but one found V2, synthesised with the combined rule and data derived formant frequencies, more natural sounding than V1, the rule based version. The one exception, an inexperienced listener, did not vote for either of the two versions.

Two inexperienced subjects did not mark a preference for either V1 or V2. One of them argued that his choice would depend on the situation in which he would be exposed to synthetic speech. He would prefer V1 in a situation demanding a high degree of intelligibility. One subject that did mark V2 as his preferred choice said that he found V3, the training version, easiest to listen to.

	V1	V2	N/A
naturalness	0	14	1
preference	1	12	2

Table 2. Votes for naturalness and preference for V1 and V2.

The sole subject that preferred the rule based version, V1, to V2 was an experienced listener. In fact, that was the male speaker of the speech database. Several comments were made on the perceived spectral contents. Many listeners commented that there were more, and sometimes disturbing high frequency components in V1. Conversely, some subjects described V2 as being darker, softer and more reduced. A couple of subjects reported perceived prosodic differences that might have been triggered by the spectral cues (formant frequency trajectories.)

Two remarks can be made regarding the results of the perceptual test. Firstly, it should be noted that the rule based system has been developed using a lower speaking rate than the one used here. It is possible that the rule based version would have obtained relatively higher voting scores, had it been synthesised with the normal segment durations.

The second remark concerns the material used for the perceptual test. All the synthesised sentences were present in corpus one which was part of the training material. Therefore, an additional regression tree analysis was made on the speech material excluding the utterances used for the perceptual test. Synthesising the test sentences using the new regression trees resulted in a version that contained some audible differences compared to V2. Judging these differences to be very small we think it is reasonable to believe that the overlap between test and training materials had an negligible impact on the results of the perceptual test.

In this study we have used the same value for *w* globally. However, this parameter could also be used, in combination with stress parameters, to dynamically vary the speech in the hyper-hypo dimension, hence adapting to locally constrained needs for clear speech.

6. CONCLUSIONS

We have implemented a data driven regression algorithm in the KTH formant based TTS system. The goal of this work is to profit from the advantages of data driven methods while maintaining the flexibility and interpretability that rule based formant synthesis provides. The current implementation allows synthesis parameters to be set to values based on weighted contribtutions from the rules and the data derived predictions. Given a database with reduced speech this means that we can continuously vary the degree of reduction since the rule derived speech resembles hyper-speech. An informal perceptual test showed that a large majority of the listeners preferred speech produced by the combined data driven and rule based approach to the rule derived speech. The former speech was also judged to be more natural than the latter.

Future development of the system will include data derived predictions of more synthesis parameters such as higher formants and voice source parameters. In particular, it will be important to obtain coherent spectral and prosodic predictions. Data driven consonant modelling should also be developed in the long term.

7. ACKNOWLEDGEMENTS

This work has been supported by grants from The Swedish National Language Technology Programme.

8. REFERENCES

[1] Moulines E. and Charpentier F. "Pitch synchronous Waveform Processing Techniques for Text-to Speech Synthesis using Diphones", *Speech Communication*, Vol 9, no 5, pp 453-467, 1990.

[2] Donovan R. and Woodland P. "Improvements in an HMM-Based Speech Synthesiser", *Proceedings of Eurospeech'95*, pp 573-576, Madrid, 1995.

[3] Carlson R., Granström B. and Hunnicutt S. "Multilingual text-to-speech development and applications", in Ainsworth A. (ed.), *Advances in speech, hearing and language processing.* London: JAI Press, UK, 1991.

[4] Breiman L., Friedman J., Olshen R., Stone C. *Classification and regression trees*, Belmont, CA: Wadsworth, 1984.

[5] Bahl L., Souza P., Gopalakrishnan P., Nahamoo D. and Picheney M. "Context dependent modeling of phones in continuous speech recognition using decision trees" *Proceedings from DARPA Speech and natural language workshop*, pp 264-269, 1991.

[6] Högberg J. "A phonetic investigation using binary regression trees", *Papers from the Eighth Swedish Phonetics Conference*, Lund, 1994.

[7] Carlson R. and Nord L. "Vowel dynamics in a text to speech system - some considerations", *Proceedings of Eurospeech'93*, pp 1911-1914, Berlin, 1993.

[8] ESPS/waves+ v5.0 manuals, Entropics Research Laboratories Inc., 1993.

[9] Sjölander K. and Högberg J. "Trying to improve phone and word recognition using finely tuned phone-like units", *PHONUM 4*, Umeå, 1997.

[10] Lindblom B. "Explaining phonetic variation: A sketch of the H&H theory", pp 403-439 in Hardcastle W. and Marchal A. (eds.) *Speech production and speech modeling*, Dordrecht: Kluwer, 1990.