Exploring data driven parametric synthesis

Rolf Carlson[1], Kjell Gustafson[1,2]
[1] KTH, CSC, Department of Speech, Music and Hearing, Stockholm, Sweden
[2] Acapela Group Sweden AB, Solna, Sweden

Abstract

This paper describes our work on building a formant synthesis system based on both rule generated and database driven methods. Three parametric synthesis systems are discussed: our traditional rule based system, a speaker adapted system, and finally a gesture system. The gesture system is a further development of the adapted system in that it includes concatenated formant gestures from a data-driven unit library. The systems are evaluated technically, comparing the formant tracks with an analysed test corpus. The gesture system results in a 25% error reduction in the formant frequencies due to the inclusion of the stored gestures. Finally, a perceptual evaluation shows a clear advantage in naturalness for the gesture system compared to both the traditional system and the speaker adapted system.

Introduction

Current speech synthesis efforts, both in research and in applications, are dominated by methods based on concatenation of spoken units. Research on speech synthesis is to a large extent focused on how to model efficient unit selection and unit concatenation and how optimal databases should be created. The traditional research efforts on formant synthesis and articulatory synthesis have been significantly reduced to a very small discipline due to the success of waveform based methods. Despite the well motivated current research path resulting in high quality output, some efforts on parametric modelling are carried out at our department. The main reasons are flexibility in speech generation and a genuine interest in the speech code. We try to combine corpus based methods with knowledge based models and to explore the best features of each of the two approaches. This report describes our progress in this synthesis work.

Parametric synthesis

Underlying articulatory gestures are not easily transformed to the acoustic domain described by a formant model, since the articulatory constraints are not directly included in a formant-based model. Traditionally, parametric speech synthesis has been based on very labour-intensive optimization work. The notion analysis by synthesis has not been explored except by manual comparisons between hand-tuned spectral slices and a reference spectrum. When increasing our ambitions to multi-lingual, multi-speaker and multi-style synthesis, it is obvious that we want to find at least semi-automatic methods to collect the necessary information, using speech and language databases. The work by Holmes and Pearce (1990) is a good example of how to speed up this process. With the help of a synthesis model, the spectra are automatically matched against analysed speech. Automatic techniques such as this will probably also play an important role in making speaker-dependent adjustments. One advantage with these methods is that the optimization is done in the same framework as that to be used in the production. The synthesizer constraints are thus already imposed in the initial state.

If we want to keep the flexibility of the formant model but reduce the need for detailed formant synthesis rules, we need to extract formant synthesis parameters directly from a labelled corpus. Already more than ten years ago at Interspeech in Australia, Mannell (1998) reported a promising effort to create a diphone library for formant synthesis. The procedure included a speaker-specific extraction of formant frequencies from a labelled database. In a sequence of papers from Utsunomiya University, Japan, automatic formant tracking has been used to generate speech synthesis of high quality using formant synthesis and an elaborate voice source (e.g. Mori et al., 2002). Hertz (2002) and Carlson and Granström (2005) report recent research efforts to combine data-driven and rule-based methods. The approaches take advantage of the fact that a unit library can better model detailed gestures than the general rules.

In a few cases we have seen a commercial interest in speech synthesis using the formant model. One motivation is the need to generate

speech using a very small footprint. Perhaps the formant synthesis will again be an important research subject because of its flexibility and also because of how the formant synthesis approach can be compressed into a limited application environment.

A combined approach for acoustic speech synthesis

The efforts to combine data-driven and rule-based methods in the KTH text-to-speech system have been pursued in several projects. In a study by Högberg (1997), formant parameters were extracted from a database and structured with the help of classification and regression trees. The synthesis rules were adjusted according to predictions from the trees. In an evaluation experiment the synthesis was tested and judged to be more natural than the original rule-based synthesis.

Sjölander (2001) expanded the method into replacing complete formant trajectories with manually extracted values, and also included consonants. According to a feasibility study, this synthesis was perceived as more natural sounding than the rule-only synthesis (Carlson et al., 2002). Sigvardson (2002) developed a generic and complete system for unit selection using regression trees, and applied it to the data-driven formant synthesis. In Öhlin & Carlson (2004) the rule system and the unit library are more clearly separated, compared to our earlier attempts. However, by keeping the rule-based model we also keep the flexibility to make modifications and the possibility to include both linguistic and extra-linguistic knowledge sources.

Figure 1 illustrates the approach in the KTH text-to-speech system. A database is used to create a unit library and the library information is mixed with the rule-driven parameters. Each unit is described by a selection of extracted synthesis parameters together with linguistic information about the unit's original context and linguistic features such as stress level. The parameters can be extracted automatically and/or edited manually.

In our traditional text-to-speech system the synthesizer is controlled by rule-generated parameters from the text-to-parameter module (Carlson et al., 1982). The parameters are represented by time and values pairs including labels and prosodic features such as duration and intonation. In the current approach some of the rule-generated parameter values are replaced by values from the unit library. The process is controlled by the unit selection module that takes into account not only parameter information but also linguistic features supplied by the text-to-parameter module. The parameters are normalized and concatenated before being sent to the GLOVE synthesizer (Carlson et al., 1991).

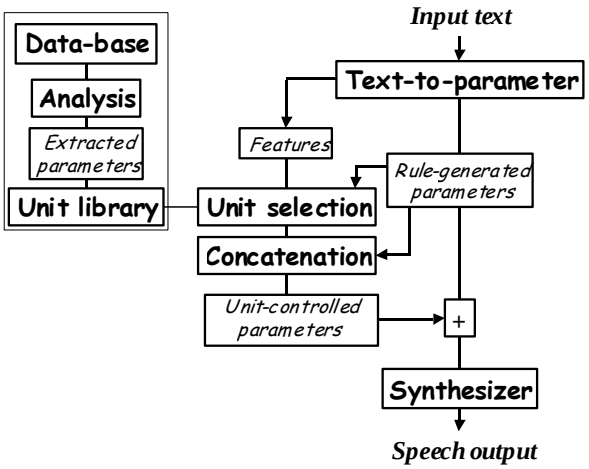

Figure 1. Rule-based synthesis system using a data-driven unit library.

Creation of a unit library

In the current experiments a male speaker recorded a set of 2055 diphones in a nonsense word context. A unit library was then created based on these recordings.

When creating a unit library of formant frequencies, automatic methods of formant extraction are of course preferred, due to the amount of data that has to be processed. However, available methods do not always perform adequately. With this in mind, an improved formant extraction algorithm, using segmentation information to lower the error rate, was developed (Öhlin & Carlson, 2004). It is akin to the algorithms described in Lee et al. (1999), Talkin (1989) and Acero (1999).

Segmentation and alignment of the waveform were first performed automatically with nAlign (Sjölander, 2003). Manual correction was required, especially on vowel–vowel transitions. The waveform is divided into (overlapping) time frames of 10 ms. At each frame, an LPC model of order 30 is created; the poles are then searched through with the Viterbi algorithm in order to find the path (i.e. the formant trajectory) with the lowest cost. The cost is defined as the weighted sum of a number of partial costs: the bandwidth cost, the frequency deviation cost, and the frequency change cost.

The bandwidth cost is equal to the bandwidth in Hertz. The frequency deviation cost is defined as the square of the distance to a given reference frequency, which is formant, speaker, and phoneme dependent. This requires the labelling of the input before the formant tracking is carried out. Finally, the frequency change cost penalizes rapid changes in formant frequencies to make sure that the extracted trajectories are smooth.

Although only the first four formants are used in the unit library, five formants are extracted. The fifth formant is then discarded. The justification for this is to ensure reasonable values for the fourth formant. The algorithm also introduces eight times over-sampling before averaging, giving a reduction of the variance of the estimated formant frequencies. After the extraction, the data is down-sampled to 100 Hz.

Synthesis Systems

Three parametric synthesis systems were explored in the experiments described below. The first was our rule-based traditional system, which has been used for many years in our group as a default parametric synthesis system. It includes rules for both prosodic and context dependent segment realizations. Several methods to create formant trajectories have been explored during the development of this system. Currently simple linear trajectories in a logarithmic domain are used to describe the formants. Slopes and target positions are controlled by the transformation rules.

The second rule system, the adapted system, was based on the traditional system and adapted to a reference speaker. This speaker was also used to develop the data-driven unit library. Default formant values for each vowel were estimated based on the unit library, and the default rules in the traditional system were changed accordingly. It is important to emphasize that it is the vowel space that was data driven and adapted to the reference speaker and not the rules for contextual variation.

Finally, the third synthesis system, the gesture system, was based on the adapted system, but includes concatenated formant gestures from the data-driven unit library. Thus, both the adapted system and the gesture system are data-driven systems with varying degree of mix between rules and data. The next section will discuss in more detail the concatenation process that we employed in our experiments.

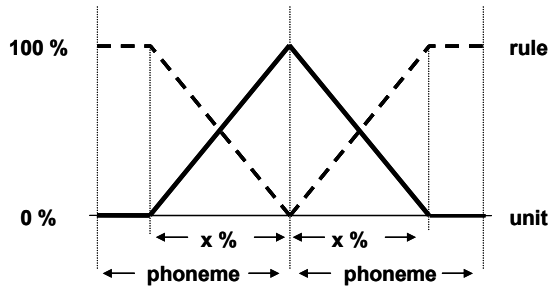

Figure 2. Mixing proportions between a unit and a rule generated parameter track. X=100% equals the phoneme duration.

Parameter concatenation

The concatenation process in the gesture system is a simple linear interpolation between the rule generated formant data and the possible joining units from the library. At the phoneme border the data is taken directly from the unit. The impact of the unit data is gradually reduced inside the phoneme. At a position X the influence of the unit has been reduced to zero (Figure 2). The X value is calculated relative to the segment duration and measured in % of the segment duration. The parameters in the middle of a segment are thus dependent on both rules and two units.

Technical evaluation

A test corpus of 313 utterances was selected to compare predicted and estimated formant data and analyse how the X position influences the difference. The utterances were collected in the IST project SpeeCon (Großkopf et al., 2002) and the speaker was the same as the reference speaker behind the unit library. As a result, the adapted system also has the same reference speaker. In total 4853 phonemes (60743 10 ms frames) including 1602 vowels (17508 frames) were used in the comparison.

A number of versions of each utterance were synthesized, using the traditional system, the adapted system and the unit system with varying values of X percent. The label files from the SpeeCon project were used to make the duration of each segment equal to the recordings. An X value of zero in the unit system will have the same formant tracks as the adapted system. Figure 3 shows the results of calculating the city-block distance between the synthesized and measured first three formants in the vowel frames.

Figure 4 presents a detailed analysis of the data for the unit system with X=70%. The first formant has an average distance of 68 Hz with a standard deviation of 43 Hz. Corresponding data for F2 is (107 Hz, 81 Hz), F3 (111 Hz, 68 Hz) and F4 (136 Hz, 67 Hz).

Clearly the adapted speaker has a quite different vowel space compared to the traditional system. Figure 5 presents the distance calculated on a phoneme by phoneme base. The corresponding standard deviations are 66 HZ, 58 Hz or 46 Hz for the three systems.

As expected, the difference between the traditional system and the adapted system is quite large. The gesture system results in about a 25% error reduction in the formant frequencies due to the inclusion of the stored gestures. However, whether this reduction corresponds to a difference in perceived quality cannot be predicted on the basis of these data. The difference between the adapted and the gesture system is quite interesting and of the same magnitude as the adaptation data. The results clearly indicate how the gesture system is able to mimic the reference speaker in more detail than the rule-based system. The high standard deviation indicates that a more detailed analysis should be performed to find the problematic cases. Since the test data as usual is hampered by errors in the formant tracking procedures we will inherently introduce an error in the comparison. In a few cases, despite our efforts, we have a problem with pole and formant number assignments.

Perceptual evaluation

A pilot test was carried out to evaluate the naturalness in the three synthesis systems: traditional, adapted and gesture. 9 subjects working in the department were asked to rank the three systems according to perceived naturalness using a graphic interface. The subjects have been exposed to parametric speech synthesis before. Three versions of twelve utterances including single words, numbers and sentences were ranked. The traditional rule-based prosodic model was used for all stimuli. In total 324=3*12*9 judgements were collected. The result of the ranking is presented in Figure 6.

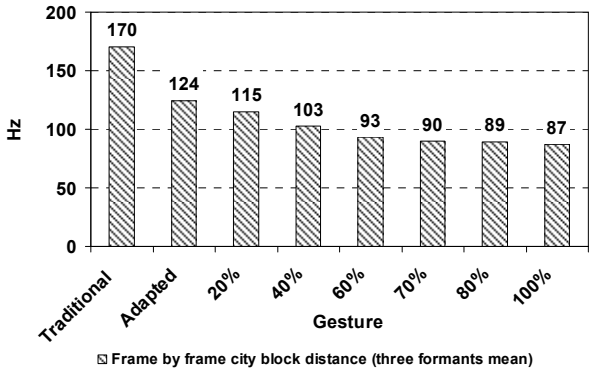

Figure 3. Comparison between synthesized and measured data (frame by frame).

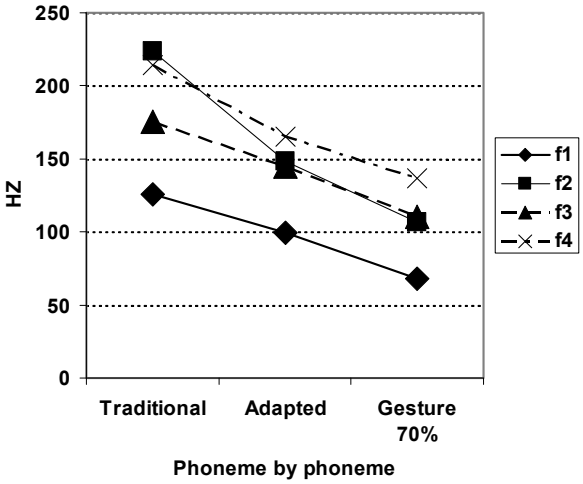

Figure 4. Comparisons between synthesized and measured data for each formant (phoneme by phoneme).

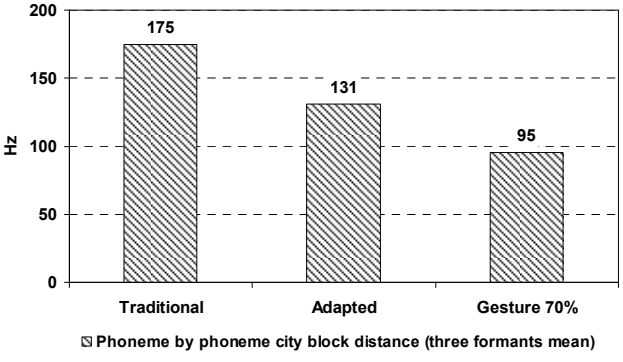

Figure 5. Comparison between synthesized and measured data (phoneme by phoneme).

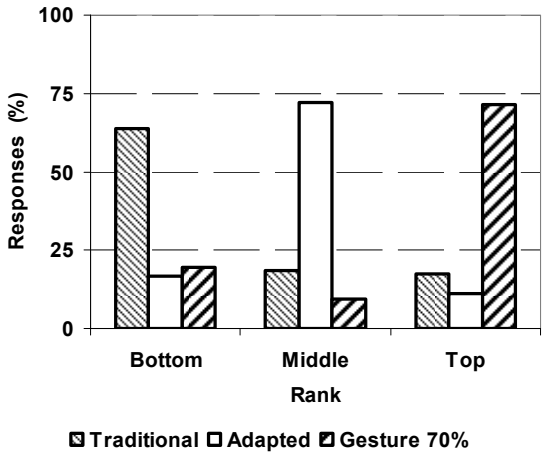

Figure 6. Rank distributions for the traditional, adapted and gesture 70% systems.

The outcome of the experiment should be considered with some caution due to the selection of the subject group. However, the results indicate that the gesture system has an advantage over the other two systems and that the adapted system is ranked higher than the traditional system. The maximum rankings are 64%, 72% and 71% for the traditional, adapted and gesture systems, respectively. Our initial hypothesis was that these systems would be ranked with the traditional system at the bottom and the gesture system at the top. This is in fact true in 58% of the cases with a standard deviation of 21%. One subject contradicted this hypothesis in only one out off 12 cases while another subject did the same in as many as 9 cases. The hypothesis was confirmed by all subjects for one utterance and by only one subject for another one.

The adapted system is based on data from the diphone unit library and was created to form a homogeneous base for combining rule-based and unit-based synthesis as smoothly as possible. It is interesting that even these first steps, creating the adapted system, are regarded to be an improvement. The diphone library has not yet been matched to the dialect of the reference speaker, and a number of diphones are missing.

Final remarks

This paper describes our work on building formant synthesis systems based on both rule-generated and database driven methods. The technical and perceptual evaluations show that this approach is a very interesting path to explore further at least in a research environment. The perceptual results showed an advantage in naturalness for the gesture system which includes both speaker adaptation and a diphone database of formant gestures, compared to both the traditional reference system and the speaker adapted system. However, it is also apparent from the synthesis quality that a lot of work still needs to be put into the automatic building of a formant unit library.

Acknowledgements

The diphone database was recorded using the WaveSurfer software. David Öhlin contributed in building the diphone database. We thank John Lindberg and Roberto Bresin for making the evaluation software available for the perceptual ranking. The SpeeCon database was made available by Kjell Elenius. We thank all subjects for their participation in the perceptual evaluation.

References

Acero, A. (1999) “Formant analysis and synthesis using hidden Markov models”, In: Proc. of Eurospeech'99, pp. 1047-1050.

Carlson, R., and Granström, B. (2005) “Data-driven multimodal synthesis”, Speech Communication, Volume 47, Issues 1-2, September-October 2005, Pages 182-193.

Carlson, R., Granström, B., and Karlsson, I. (1991) “Experiments with voice modelling in speech synthesis”, Speech Communication, 10, 481-489.

Carlson, R., Granström, B., Hunnicutt, S. (1982) “A multi-language text-to-speech module”, In: Proc. of the 7th International Conference on Acoustics, Speech, and Signal Processing (ICASSP’82), Paris, France, vol. 3, pp. 1604-1607.

Carlson, R., Sigvardson, T., Sjölander, A. (2002) “Data-driven formant synthesis”, In: Proc. of Fonetik 2002, Stockholm, Sweden, STL-QPSR 44, pp. 69-72

Großkopf, B., Marasek, K., v. d. Heuvel, H., Diehl, F., Kiessling, A. (2002) “SpeeCon - speech data for consumer devices: Database specification and validation”, Proc. LREC.

Hertz, S. (2002) “Integration of Rule-Based Formant Synthesis and Waveform Concatenation: A Hybrid Approach to Text-to-Speech Synthesis”, In: Proc. IEEE 2002 Workshop on Speech Synthesis, 11-13, September 2002 Santa Monica, USA.

Högberg, J. (1997) “Data driven formant synthesis”, In: Proc. of Eurospeech 97.

Holmes, W. J. and Pearce, D. J. B. (1990) “Automatic derivation of segment models for synthesis-by-rule”, Proc ESCA Workshop on Speech Synthesis, Autrans, France.

Lee, M., van Santen, J., Möbius, B., Olive, J. (1999) “Formant Tracking Using Segmental Phonemic Information”, In: Proc.of Eurospeech ‘99, Vol. 6, pp. 2789–2792.

Mannell, R. H. (1998) “Formant diphone parameter extraction utilising a labeled single speaker database”, In: Proc. of ICSLP 98.

Mori, H., Ohtsuka, T., Kasuya, H. (2002) “A data-driven approach to source-formant type text-to-speech system”, In ICSLP-2002, pp. 2365-2368.

Öhlin, D. (2004) “Formant Extraction for Data-driven Formant Synthesis”, (in Swedish). Master Thesis, TMH, KTH, Stockholm.

Öhlin, D., Carlson, R. (2004) “Data-driven formant synthesis”, In: Proc. Fonetik 2004 pp. 160-163.

Sigvardson, T. (2002) “Data-driven Methods for Paameter Synthesis – Description of a System and Experiments with CART-Analysis”, (in Swedish). Master Thesis, TMH, KTH, Stockholm, Sweden.

Sjölander, A. (2001) “Data-driven Formant Synthesis” (in Swedish). Master Thesis, TMH, KTH, Stockholm.

Sjölander, K. (2003) “An HMM-based System for Automatic Segmentation and Alignment of Speech”, In: Proc. of Fonetik 2003, Umeå Universitet, Umeå, Sweden, pp. 93–96.

Talkin, D. (1989) “Looking at Speech”, In: Speech Technology, No 4, April/May 1989, pp. 74–77.