Data-driven formant synthesis

David Öhlin and Rolf Carlson
CTT, Department of Speech, Music and Hearing, KTH

Abstract

A new method of speech synthesis, which combines earlier work on data-driven formant synthesis with improved data extraction and concatenation of recorded voiceless segments, has been developed and implemented as a TTS system. A listening test has been carried out, which has shown that this hybrid synthesis significantly raises the perception of naturalness in the synthesized speech compared to rule only formant synthesis.

Introduction

Current speech synthesis efforts, both in research and in applications, are dominated by methods based on concatenation of spoken units. In this paper we report on a project where rule-based formant synthesis is combined with both data-driven methods and concatenative synthesis. The paper is a continuation of the work earlier reported in Carlson et al. (2002).

Concatenative synthesis

In the review by Klatt (1987) some of the early efforts on synthesis based on concatenative synthesis are included as an alternative to rule-based formant synthesis (Carlson and Granström, 1976, Carlson et al., 1982 and Klatt, 1982). Charpentier and Stella (1986) opened a new path towards speech synthesis based on waveform concatenation, by introducing the PSOLA model for manipulating pre-recorded waveforms. The current methods using unit selection from large corpora rather than using a fixed unit inventory tries to reduce the number of units in each utterance and to solve context dependencies over a longer time frame. Möbius (2000) gives an extensive review of corpus-based synthesis methods.

Formant synthesis

The need to synthesize different voices and voice characteristics and to model emotive speech has kept research on formant synthesis active (Carlson et al., 1991). The motivation is that rule-based formant synthesis has the needed flexibility to model both linguistic and extra linguistic processes.

Thus, one can predict that formant synthesis will again be an important subject because of its flexibility and also because of how the formant synthesis approach can be squeezed into a limited application environment.

Data-driven formant synthesis

Research efforts to combine data-driven and rule-based methods in the KTH text-to-speech (TTS) system has been pursued in several projects. In a study by Högberg (1997) formant parameters were extracted from a data-base 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 this 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 more generic and complete system for unit selection using regression trees, and applied it to the data-driven formant synthesis. This system is described in Figure 1.

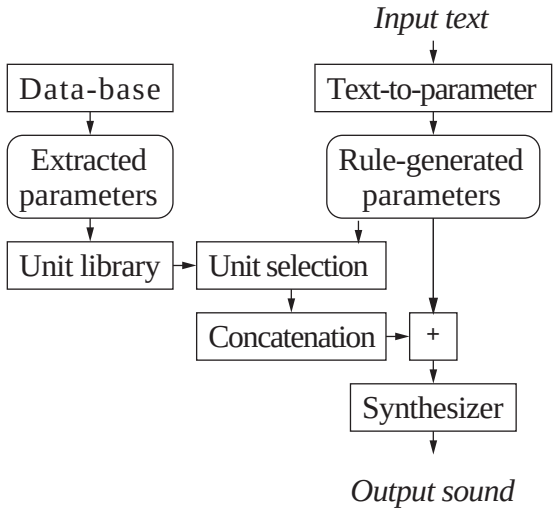

Figure 1. Data-driven formant synthesis, using a library of extracted parameters.

This approach to formant synthesis takes advantage of the fact that a unit library can better model detailed gestures than the current general rules. By keeping the rule-based model, the flexibility to make modifications and the possibility to include both linguistic and extra-linguistic knowledge sources are kept.

Formant extraction

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 (Sigvardson, 2002). With this in mind, an improved formant extraction algorithm, using segmentation information to lower the error rate, was developed (Öhlin, 2004). It is akin to the algorithm described in Lee et al. (1999), and Talkin (1989).

In the new algorithm, the recorded sound is divided into (possibly overlapping) time frames of 10 ms. At each frame, an LPC model of order 30 is created by means of the Yule-Walker equations. The frequencies and bandwidths of the poles are extracted from the roots r_n of the corresponding polynomial as:

$$f_n = f_s \cdot \arg(r_n)/2\pi \quad (1)$$

$$B_n = f_s \cdot \log(|r_n|)/\pi \quad (2)$$

All poles are then searched through with the Viterbi algorithm (Forney, 1973) 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 norm frequency, which is formant, speaker, and phoneme dependent. This requires labeling of the input before the formant tracking is carried out. Finally, the frequency change cost penalizes rapid changes in formant frequencies, and makes 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 (F5) is then discarded. The justification for this is to ensure reasonable values for the fourth formant. Furthermore, one sub-F1 pole is extracted to avoid confusion with F1. The algorithm also introduces eight times oversampling before averaging, giving a reduction of the variance of the estimated formant frequencies. After the extraction, the data is downsampled to 100 Hz.

Concatenation of voiceless consonants

Although reasonably high in intelligibility and clarity, the voiceless fricatives and plosives generated by the formant synthesizer lack in naturalness. One way of remedying this is to generate these phonemes by replacing the synthesized waveform with recorded sounds after time adjustment. This was implemented and integrated into the KTH test-to-speech system, and a listening test was carried out, which showed that this method—by itself—led to a raise in the perception of naturalness, but also that the raise is even bigger when combined with data-driven formant synthesis (Vinet, 2004).

In the synthesis, unvoiced phones are taken from a database of recorded diphones. Each synthesized fricative is replaced by the fricative portion of the two corresponding diphones. The recorded sounds are time-scaled using TD-PSOLA (Charpentier and Moulines, 1990) and concatenated with the formant synthesized wave-form.

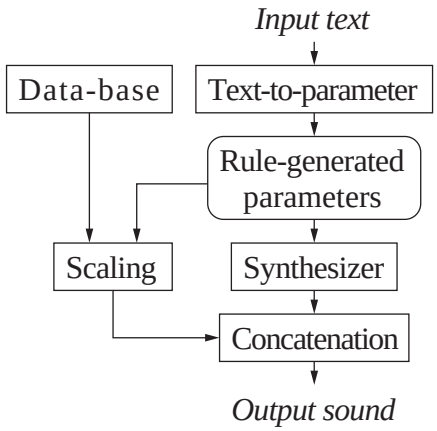

Figure 2. Scaling and concatenation of unvoiced consonants with formant synthesized speech.

Since the affected phonemes are voiceless, no pitch needs to be considered when scaling and concatenating, making the task significantly easier. The phoneme /h/, however, is often voiced in natural speech, and is therefore not included in the concatenative synthesis, even though the formant synthesizer treats is as a voiceless phoneme.

For the unvoiced plosives, only the release phase (the burst) is included. Since the release is just a short-term signal, it needs not be scaled to fit the synthesized speech.

The TTS system

The data-driven synthesis and the concatenation of voiceless consonants were combined and integrated into a full TTS system. Since the system as a whole relies heavily on a proper driving database, special care had to taken to ensure that the segmentation of the diphone data-base was reasonable, and that the extracted formant trajectories were without errors.

Segmentation and alignment were first performed automatically with nAlign (Sjölander 2003). Manual correction was required, especially on vowel–vowel transitions. With a reasonable segmentation, the formants were extracted from all the files. Special tools were constructed in order to automatically detect errors in the extracted data, for instance by comparing the data with formant data that had been generated by rules, or detecting rapid frequency changes that had slipped through.

The current TTS system is displayed in Figure 3. First, the input is transcribed and converted, by rules, into parameter data. The rule-generated formant parameters are replaced by data from the unit library, which is properly time scaled and concatenated. The sound wave is then synthesized using the GLOVE synthesizer (Carlson et al. 1990), and the unvoiced consonants are included using the PSOLA technique described above.

Some corrections had to be included when building and testing the system. One problem still remaining to handle is the realization of /h/. The contextual dependency of the phoneme is not very well modeled by the diphone approach. Thus an alternative method needs to be developed. Currently a simple linear interpolation between the neighboring phones is used as a back off model. However, it need to be refined to get the same quality as the rest of the system.

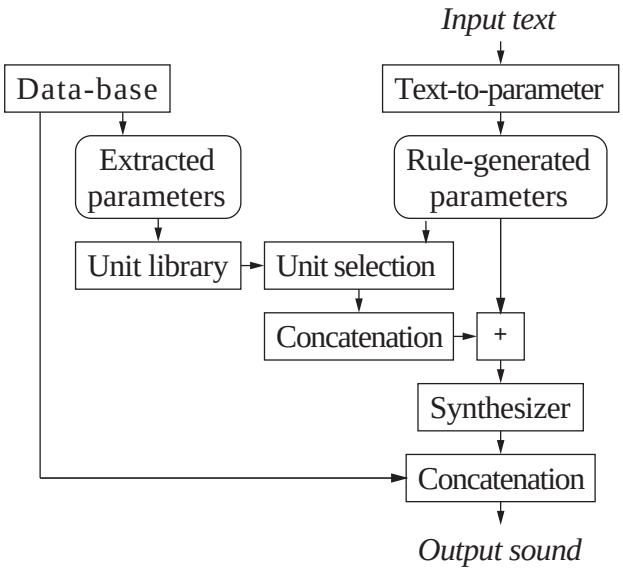

Figure 3. The proposed data-driven TTS system, described in detail in the text.

Evaluation

A preliminary listening test was carried out to evaluate the combined synthesis. 12 test subjects were asked to listen to 10 different sentences, all of which had been synthesized in two different ways: by rule, and by data. For each such pair, the test subjects were asked to pick the one that was most natural sounding.

The data-driven synthesis was considered more natural 73 % of the time. On average, none of the sentences was perceived as more natural when synthesized with rules only, than with the new synthesis.

The conclusion of this and the earlier listening tests is that the data-driven formant synthesis is in fact perceived as more natural sounding than the rule-driven synthesis. However, it is also apparent that a lot of work still needs to be put into the automatic building of a formant unit library.

Acknowledgements

This research was carried out at the Centre for Speech Technology, a competence centre at KTH, supported by VINNOVA (The Swedish Agency for Innovation Systems), KTH and participating Swedish companies and organizations.

References

Carlson R., Granström B. (1976) A text-to-speech system based entirely on rules. In: Proceedings of ICASSP 76.

Carlson R., Granström B., Hunnicutt S. (1982) A multi-language text-to-speech module. In: Proceedings of ICASSP 82. pp. 1604–1607.

Carlson R., Granström B., Karlsson I. (1991) Experiments with voice modelling in speech synthesis. In: Speech communication, No. 10, pp. 481–489.

Carlson R., Sigvardson T., Sjölander A. (2002) Data-driven formant synthesis. In: Fonetik 2002.

Charpentier F., Stella M. (1986) Diphone Synthesis Using an Overlap-Add Technique for Speech Waveforms Concatenation. In: Proceedings of ICASSP 86 (3), pp. 2015–2018.

Charpentier F., Moulines E. (1990) Pitch-Synchronous Waveform Processing Techniques for Text-to-Speech Synthesis Using Diphones. In: Speech Communication, Vol. 9, No 5/6, Dec. 1990, pp. 435–467.

Forney Jr G. D. (1972) The Viterbi Algorithm. In: Proceedings of the IEEE, Vol. 61, No 3, March 1973, pp. 268–277.

Högberg J. (1997) Data driven formant synthesis. In: Proceedings of Eurospeech '97, pp. 565–568.

Klatt D. (1982) The Klattalk Text-to-Speech Conversion System. In: Proceedings of ICASSP 82, pp. 1589–1592.

Klatt D. (1987) Review of Text-to-Speech Conversion for English. In: Journal of the Acoustical Society of America, Vol. 82, No 3–87, pp. 737–793.

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

Möbius B. (2000) Corpus-based speech synthesis: methods and challenges. In: Arbeitspapiere des Instituts für Maschinelle Sprachverarbeitung (Univ. Stuttgart), AIMS 6 (4), pp. 87–116.

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

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: Proceedings of Fonetik 2003, Umeå Universitet, Umeå, pp. 93–96.

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

Vinet R. (2004) Enhancing Rule-Based Synthesizer Using Concatenative Synthesis. Master Thesis, TMH, KTH, Stockholm.

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