Data-driven formant synthesis

Carlson, R.

journal: Proceedings of Fonetik, TMH-QPSR
volume: 44
number: 1
year: 2002
pages: 121-124

Data-driven formant synthesis

Rolf Carlson, Tor Sigvardson and Arvid Sjölander
CTT, Department of Speech, Music and Hearing, KTH

Abstract

A new approach to formant synthesis, using both rule-based and data-driven methods is presented. Preliminary results from a feasibility study show an advantage in speech quality compared to a rule-based reference system. The method is now being implemented into a general framework.

Introduction

Current speech synthesis efforts, both in research and in applications, are dominated by methods based on concatenation of spoken units. In some cases the original waveform is simply used as it is, but often it is processed to some degree before use. Research on speech synthesis is to a large extent focused on how to model efficient unit selection and unit concatenation and how optimal data-bases should be created. The research efforts on formant synthesis or articulatory synthesis have been reduced. In this paper we report on a project where rule-based formant synthesis is combined with data-driven methods.

Concatenative synthesis

In the review by Klatt (1987) some of the early efforts on synthesis based on concatenative synthesis are included. Already Peterson et al. (1958) suggested that unit concatenation might be a possible solution for speech synthesis. Dixon and Maxey (1968) made a special effort to create a unit library for diphone synthesis. Early synthesis research at AT&T based on "Diadic Units" (Olive, 1977) demonstrated 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. However, the flexibility is also a problem, since for example articulatory constraints are not included in the model. The underlying articulatory gestures are not easily transformed to the acoustic domain described by the formant model. However, successful efforts to go "halfway" using Higher-level articulatory based parameters have been reported by Stevens and Bickley (1991) and Ogden et al. (2000).

An alternative approach to reduce the need for detailed formant synthesis rules, but still keeping the flexibility of the formant model, is to extract formant synthesis parameters directly from a labelled corpus. Mannell (1998) has reported a promising effort to create a diphone library for formant synthesis based on a data-base.

Research efforts to combine data-driven and rule-based methods in the KTH text-to-speech 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.

Recently we have seen a renewed commercial interest for speech synthesis using the formant model (e.g. Aurix TTS from 20/20 Speech). One motivation is the need to generate speech, using very small "foot print".

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.

A combined approach

This paper describes a new effort to combine a rule-based formant synthesis approach and a corpus-driven approach. The approach takes advantage of the fact that a unit library can better model detailed gestures then the current general rules. In the current work the rule system and the unit library is more clearly separated compared to our earlier work (e.g. Högberg, 1997). However, by keeping the rule-based model we also keep the flexibility to make modifications and the possibility to include both linguistic and extralinguistic knowledge sources.

Figure 1 illustrates the approach from a technical point of view. A data-base is used to create a unit library. 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.

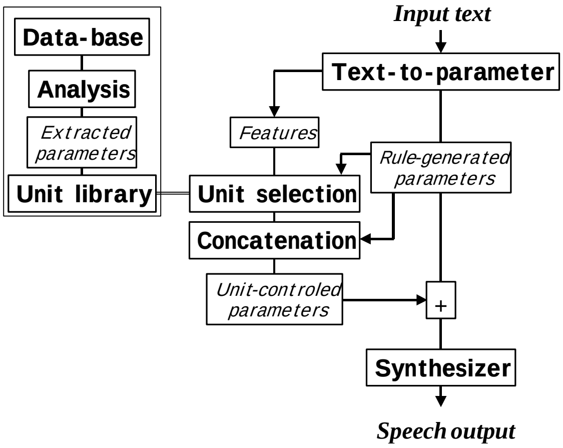

Figure 1. Rule-based synthesis system using a data-driven unit library.

In our traditional text-to-speech system the synthesiser 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 synthesizer.

Pilot experiment

A feasibility study was carried out to evaluate the data-driven formant synthesis approach for Swedish (Sjölander, 2001). The synthesis quality from three systems using different methods to create the unit library was evaluated concerning "speech clarity" and "naturalness". In this preliminary evaluation, no linguistic features besides the phoneme labels were used in the unit selection module. The systems were also compared to a rule-based system as reference.

Evaluated systems

Reference system

As reference the KTH rule-based formant synthesis system was used.

Rule-based unit system

For comparison a diphone unit library was created using the *Reference system* (Figure 2). The same diphone inventory was used as in the data-based unit systems, described below. The first four formants in the selected units replaced the rule generated ones, after duration adjustments and smoothing in the concatenation module. The unit-controlled and the rule-generated parameters were then used as input to the synthesizer.

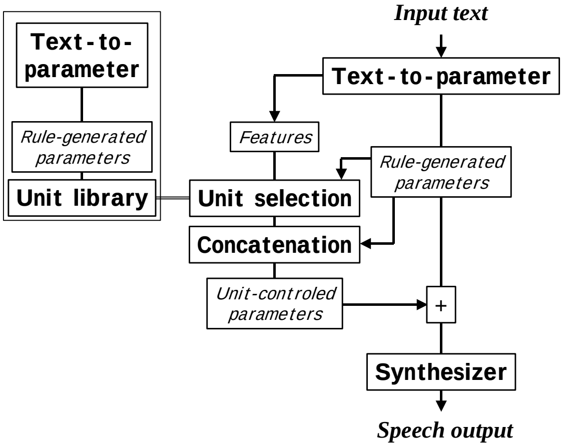

Figure 2. Rule-based synthesis system using a rule-generated unit library.

Data-based unit system

In this experiment we were able to use the “Ingemar corpus” recorded and labelled by Babel Infovox for one of their MBROLA based products. The first four formants were measured for each diphone using the Waves formant tracking software and included in the unit library according to Figure 1. Only the diphones that were needed for synthesis of the test utterances were manually corrected using the Wavesurfer software. The first four formants in the selected units replaced, after duration adjustments and smoothing, the rule generated ones and were used as input to the synthesizer.

V data-based unit system

In the *Data-based unit system* the formant extraction was not accurate enough, even after some manual correction. Therefore, a mixed system *V data-based unit system* was developed where only the vowel formants were extracted from the data-base, while the *Rule-based unit system* was used to generate the consonantal parts in the unit library.

Test material

Four Swedish utterances were used in the experiment:

”Erfarenhet är det namn alla ger sina misstag.”
”Kräftor kräva dessa drycker.”
”Den oskicklige smeden klandrar järnet.”
”Avundsamma ögon blir aldrig mätta.”

Test procedure

Each utterance was synthesized by the four systems. 13 subjects were asked to make a comparison of the synthesized output from each system and rank the relative “clarity” or the relative “naturalness” in separate sessions. Each utterance was graphically presented on the computer screen by four buttons, corresponding to the four synthesis systems. By pressing one button the subject could listen to one synthesized version of the utterance. The subjects could repeat a stimuli as many times as they wanted. The four utterances was presented twice in random order in one session giving 26 observations on the clarity ranking and in a separate session 26 observations on the naturalness ranking. The subjects had no detailed knowledge about speech synthesis but had been exposed to synthesis before the experiment was carried out.

Results

The results from the pilot experiment are presented in Table 1 and Table 2. The data, in percent, show wheather the systems labelled on each row are ranked by the subjects to have preference over the systems labelled in each column. For example, in Table 1 the *V data-based unit system* is regarded to be better compared to the *Reference system* in 60 percent of the observations, while only 29 percent of the observations gave preference to the *Rule-based unit system* compared to the *Reference system*.

The results in Table 1 show that the *V data-based unit system* was judged to give the best clarity. It is judged better than all other systems. The transformation of the *Reference system* to the *Rule-based unit system* gave a surprisingly large quality reduction, indicating that the *Reference system* captures the transitions better than the primitive concatenation method that was used in the experiment. However, if the vowel transitions are extracted from the data-base, the output is preferred despite the fact that the formant parameter tracks are concatenated. The consonant gestures are not captured well in the *Data-based unit system*. One can suspect that this is caused by less accurate measurements, rather than the method itself. This was also the reason for the creation of the mixed system.

V data-based u.	60	73	64
Data-based u.	38	45	-
Rule-based u.	29	-	-
Reference	-	-	-
	Reference	Rule-based u.	Data-based u.

Table 1. Clarity preference in percent.

V data-based u.	73	89	49
Data-based u.	68	84	-
Rule-based u.	24	-	-
Reference	-	-	-
	Reference	Rule-based u.	Data-based u.

Table 2. Naturalness preference in percent.

The *V data-based unit system* method was also the preferred system according to the ranking in the naturalness dimension. It is interesting to note that even the method using data-based consonant transitions (*Data-based unit system*) was preferred compared to the rule generated systems and was actually regarded to be equal to the *V data-based unit system.* The result suggests that the vocalic parts in the synthesis are more important for the naturalness than the consonantal parts, while both the vocalic and the consonantal parts are of equal importance for the clarity.

Conclusion

We have in this paper presented a new approach building formant-synthesis systems based on both rule-generated and data-base driven methods. A pilot experiment was reported showing that this approach can be a very interesting path to explore further. Despite a very simple implementation the preliminary results showed an advantage in naturalness compared to the traditional reference system. Work is currently on the way to create a generic platform to continue this research on formant synthesis methods, based on both rules and unit-concatenation.

Acknowledgements

We thank Babel Infovox for making the Ingemar data-base available for the experiment. Kåre Sjölander gave valuable support in several critical parts of the project. 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, and Granström B (1976). A text-to-speech system based entirely on rules. In*: Proceedings ICASSP-76.*

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

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

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

Dixon N R, and Maxey H D (1968). Terminal Analog Synthesis of Continuous Speech Using the Diphone Method of Segment Assembly. *IEEE Trans. AudioElectroacoust.* AU-16, 40-50.

Högberg J (1997). Data driven formant synthesis. In: *Proceedings of Eurospeech 97*.

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

Klatt D (1987). Review of Text-to-Speech Conversion for English. *Journal of the Acoustical Society of America*. 82: (3) 737-793.

Mannell R H (1998). Formant diphone parameter extraction utilising a labeled singlespeaker database. In: *Proceedings of ICSLP 98*.

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

Ogden R, Hawkins S, House J, Huckvale M, Local J, Carter P, Dankovicova J, and Heid S (2000). Prosynth: An Integrated Prosodic Approach to Device-Independent, Natural-Sounding Speech Synthesis. *Computer Speech and Language.* 14: 177-210.

Olive J P (1977). Rule synthesis of Speech from Diadic Units. In: *Proceedings of ICASSP-77*. 568-570.

Peterson G, Wang W, and Sivertsen El (1958). Segmentation Techniques in Speech Synthesis, *Journal of the Acoustical Society of America.* 32: 639-703.

Sjölander A (2001). Datadriven formantsyntes, master thesis TMH, KTH, Stockholm (in Swedish). http://www.speech.kth.se/ctt/publications .

Stevens K N, and Bickley C A (1991). Constraints among parameters simplify control of Klatt formant synthesizer, *Journal of Phonetics 19*, 161-174.