Evaluation of Perceptual Quality of Control Point Reduction in Rule-Based Synthesis

Kimmo Pärssinen, Marko Moberg

Nokia Technology Platforms
Tampere, Finland
{Kimmo.Parssinen, Marko.Moberg}@nokia.com

Abstract

Text-to-speech implementations on embedded devices usually require low memory consumption and computational complexity. Due to its simplicity, formant synthesizer is still an attractive solution for some applications. The formant values and transitions are controlled by a set of rules, which assign control points for synthesis parameters. This paper investigates the possibility to reduce the number of control points for formant contours from four to two per phoneme. The reduced model contained only the values at the end of the onset transition and in the beginning of the offset transition. Various interpolation techniques were studied but linear interpolation was used for its simplicity. The 4- and 2-point models were compared in a listening evaluation test. The results show that the reduction of control points does not have any effect on the perceived quality. The dynamic, context dependent positioning of the two control points preserves the most essential information of formant contours.

Index Terms: speech synthesis, speech perception, formant contour

1. Introduction

Text-to-speech (TTS) systems for embedded devices must comply with the constraints set by the available memory size and the processing power. On the other hand, the speech quality should not be overly compromised while pursuing the smaller size and the lighter computational complexity. One of the most attractive engine solutions for very low footprint synthesis is still formant synthesizer developed by Klatt [1][2]. It provides adequate speech quality for many applications requiring only short utterances while consuming very little memory. The engine itself needs only about 30 kilobytes. The implementation of the Klatt-based synthesizer used in this paper has been introduced in several Nokia N-series and E-series phones as a feedback mechanism for speaker independent name dialing feature [3].

The speech quality of the Klatt-based TTS system is mainly determined by the implementation of the rules controlling the Klatt parameters. The rules should provide the parameter values for each phoneme (e.g. formants, formant bandwidths) and also model the coarticulation effects through proper parameter transitions. The formant changes from one phoneme to another may be modeled by defining a type (e.g. discontinuous, smooth) and length of the transition [4][5]. The precise mimicking of formant contours of human speech would provide an ideal solution for synthesis but it would require complex calculations such as solving second-degree differential equations [6]. Another approach would be to use a large number of templates for different contexts but that would consume large amounts of memory. Instead, the practical solution would be to approximate the contours by defining certain control points and then interpolating the values between the points.

It has been shown in the past research that a coarse representation of formant contours for vowels e.g. using 20% and 80% points is adequate for their correct identification [7]. The increase in the modeling complexity does not necessarily improve the identification accuracy. The best vowel identification rate is obtained by determining the formant contours based on the onset, target and the offset values of the formants [8]. Based on these results we developed a framework for defining formant contours for each phoneme using four control points. To further reduce the complexity, we made the simplified version of the framework using only two control points. This paper compares differences in perceptual quality between the two alternative sets of control points for formant contours. The study is not limited to vowels but the same control point framework is applied to other type of phonemes as well. The formant transitions are studied in their normal word context and not in isolation. Furthermore, the perceptual impact of various interpolation techniques is discussed.

2. Synthesis framework

The TTS system used in this study consists of a completely language-independent engine and a language-specific data that is loaded into memory during synthesis [9]. The top-level system diagram showing the main functional modules is presented in Figure 1.

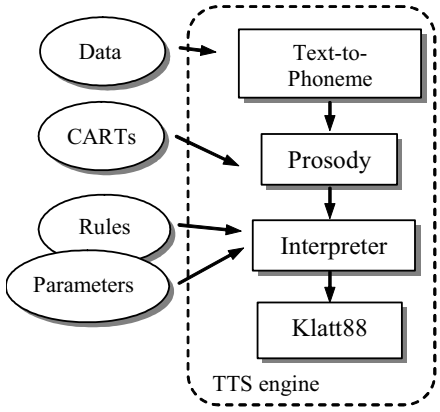

Figure 1: Block diagram of the TTS system.

The TTS engine includes the following data configurable and language-independent modules: The Text-to-Phoneme (TTP) module, a prosody module, a rule processing module and an

actual formant synthesizer (Klatt88). The language data consists of e.g. lookup-tables for TTP, prosody parameters for prosody creation, and phoneme parameters and language-specific rules for controlling the parameters. The prosody module uses CARTs (Classification And Regression Trees), which have been automatically generated from annotated speech data and encoded into a binary format [10][11]. The rules and the parameters were developed using recorded reference utterances and literature [12][13]. This paper concentrates on the rules and the framework for controlling the synthesis parameters. The main focus is on the formant frequencies and on the way their behavior is modeled.

2.1. Representation of formant contours

The rule framework allows the control of each synthesis parameter. The rules are context dependent taking into account e.g. the manner of the neighboring phonemes and they are applied to one phoneme at a time. They may determine initial value of a parameter at the phoneme boundary, the duration of the onset transition, the duration of the offset transition, and the final value of the parameter at the end of the phoneme. An example of the four control points for formants and the variables, which are used in determining them, are shown in Figure 2. The abovementioned four control points were used in optimizing the rules in our reference implementation.

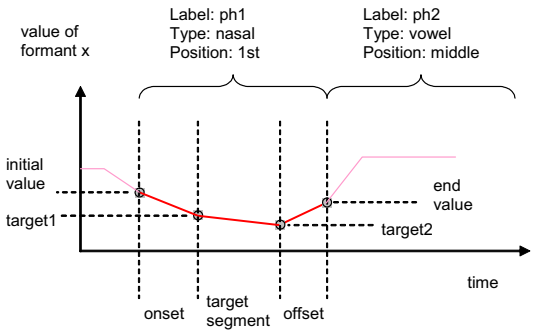

Figure 2. The available control variables for formant x of a phoneme ph1.

The initial value may the last value of the previous phoneme (to provide a smooth transition) or another value set by the rules (discontinuous transition). The duration of the onset and the offset segments are often (i.e. for sonorants) proportional to the length of the phoneme. In addition to the duration of the offset segment the steepness of the transition is controlled by the end value. It may be set to reach e.g. 20% of the target value of the following phoneme so that most of the transition (80%) will take place during the onset period of that phoneme. There are also additional controls for diphthongized vowels. Another set of target frequencies (target2) for F1, F2 and F3 are provided to set different values for the end of the target segment.

Encouraged by the results in [7] and [8], we wanted to find out if the reduction of the control points affects the perceived quality of the synthetic utterances. The number of control points for formants F1-F4 was reduced from four to two. Instead of using fixed measures of formant contours at e.g. 20% and 80% points of the phoneme duration, we used the same context dependent rules as in the 4-point implementation in assigning the location of control points. The two points defined by the rules were 1) the end of the onset segment, and 2) the beginning of the offset segment as shown in Table 1.

Table 1. Definition of two control points for parameter contours.

Control points		Defined by	
1	End of onset segment	% from the beginning of the phoneme	Target value 1
2	Beginning of offset segment	% from the end of the phoneme	Target value 2

The rules individually assigned the durations of onset and offset segments for all the formants from F1 to F4. The actual formant values, which correspond to the "steady-state" of the phoneme, were taken from a Klatt parameter table. The values in the table were obtained from natural speech at the earlier stages of the system development. The values were the same for the both points except in the case of diphthongs where another set of values for F1, F2 and F3 was applied to the second point. The formant contour using two control points is illustrated in Figure 3.

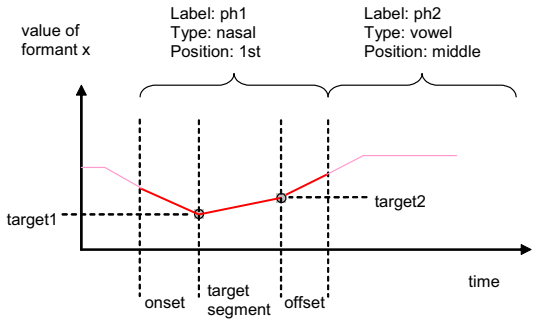

Figure 3. Two control points for formant x of a phoneme ph1.

The 2-point model lacks the control over the shape of the transitions from and to the target segment end points (target1 and target2). The shapes of the transitions are determined purely by the interpolation scheme used. In the case of linear interpolation (as shown in the figure), the start and end points of the target segment are connected to previous and following phonemes by straight lines.

2.2. Interpolation techniques

Two different interpolation techniques and their variations were experimented in this work. The interpolation was done for each phoneme having either two or four fixed control points. The first and the simplest method was the linear interpolation between data points. The clear advantages of linear interpolation are simplicity and low computational complexity. The disadvantage is that it creates a piecewise contour and the resulting formant transitions are not smooth, especially when the control points are sparse. Another interpolation method applied was the natural cubic spline

interpolation. A spline function consists of polynomial pieces on sub-intervals, i.e. between control points, joined together with certain continuity conditions. A natural cubic spline will result in a smoothest possible interpolating function given the control points. However, natural cubic spline functions are not directly applicable when the application requires that there are no over or under shoots between the given control points. The overshoots and oscillations may be avoided using so called tension splines. When an additional parameter tension τ is given a large value, the curve passing through the data points will have high tension. This can be interpreted as a force that stretches the curve tightly among the given control points and therefore there are no over or under shoots. The resulting curve approaches the piecewise linear function (a spline of degree 1) when $\tau \rightarrow \infty$. Since direct solving of tension splines would lead to hyperbolic functions, they were not applied as such in this study but smoothing of the piecewise linear interpolation was experimented instead. In this approach running mean function was run over the interpolated data points. This results in a curve close to tension splines. [14]

2.3. Application of method

Several test utterances were synthesized using two and four control points with various interpolation techniques. An example of the differences between the two-point and four-point models with linear interpolation is shown in Figure 4. The formant contours in the graph represent the three phonemes /f/, /aI/ and /v/ (SAMPA notation) of the word "five".

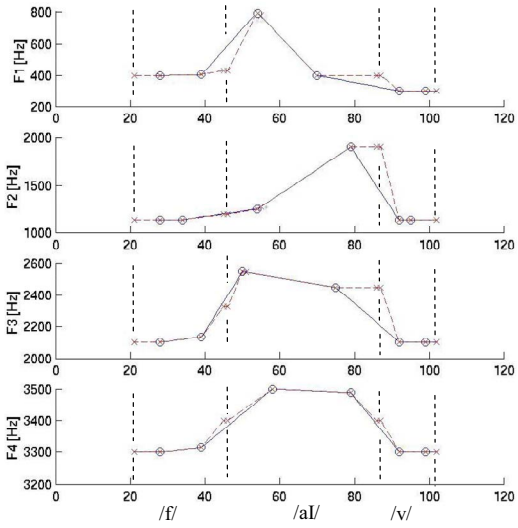

Figure 4. Formant contours (F1-F4) of two and four point linear interpolation of the utterance "five".

The dashed line represents the contour interpolated using four points per phoneme. The actual points are marked with 'x'. The solid line with 'o' marks denotes the contour interpolated using two points. The two mid-points of each phoneme are the same in both cases. The x-axis shows the time in 5 ms synthesis frames e.g. 40 corresponds to 200 ms.

The only differences between the two versions are located in transitions in phoneme boundaries. The visible details of the transitions in the four-point model can not be reproduced in the same way with the two-point version. It should be noted that the same 4- and 2-point modeling was applied to all the phonemes and not just to vowels.

3. Evaluation

The effect of the control point reduction was assessed with a listening evaluation test. The 4-point and 2-point formant contour models with linear interpolation were selected to the test. Cubic spline interpolation and linear interpolation with smoothing were also applied to some utterances but they were not included in the listening evaluation. Instead, a subjective evaluation was made.

3.1. Participants

The total of eight listeners took part in the listening evaluation test. Six of the participants were male and two of them were female. All the listeners were between 25 and 40 years of age, non-native English speakers with fluent skills in English, and with previous experience of synthetic speech.

3.2. Test data

Two different interpolation strategies were applied to 23 synthetic utterances. The sampling frequency of the synthesis output was 10 kHz but all the utterances were up-sampled to 16 kHz. A single channel (mono) signal with 16 bits per sample was used.

The utterances were isolated words or word pairs (first name and last name), which were selected to get the full coverage of 44 US English phonemes. The test data was chosen to get a good overview of the synthetic quality rather then detailed knowledge of any particular phoneme transitions.

3.3. Test procedure

The synthesized utterances were rated using comparative MOS (mean opinion score) listening evaluation test. The test was carried out in an uncontrolled manner where each participant listened to and rated the utterances on their own PC using dedicated listening evaluation test software and headphones. Two utterances were played back to back and participants had to rate how the second utterance compared with the first one. The rating was made using a five point scale (-2 = worse, -1 = slightly worse, 0 = about the same, 1 = slightly better, 2 = better).

4. Results

The results of the listening evaluation test revealed that there were no differences in perceived quality when 2-point and 4-point utterances were compared. The mean CMOS score of the test utterances was -0.01 with confidence interval between -0.05 and 0.04. In other words there was no significant preference.

The subjective analysis of the other interpolation methods was also performed. The cubic spline interpolation produced strong over and undershoots, which were perceived partly as over articulation and partly as speech quality degradations. The linear interpolation with smoothing, equivalent of tension

splines, produced smooth contours but there was no noticeable improvement in synthetic speech quality.

The use of two control points instead of four reduced the number of slope calculations of linear interpolation by two. Because the rules controlling the start and end points were made redundant, the memory footprint of the language-specific rules was reduced by 13%.

5. Discussion

The coarse 2-point representation of the formant contours in formant synthesis seems to be adequate in terms of perceived quality. Although the differences between the 4-point model and 2-point model are clear in visual inspection, they can not be heard in the actual word context.

The information that is lost during the reduction of the control points is affecting the shape of the transitions at onset and offset of a phoneme. Such transitions occurring e.g. in plosive-vowel boundaries are usually perceptually important. The fewer control points can still produce the contour shape that is perceptually indistinguishable from the 4-point reference. One of the reasons is the insensitivity of human auditory system to certain types of variations in transition shape. Another reason is illustrated in the sketched example in Figure 5.

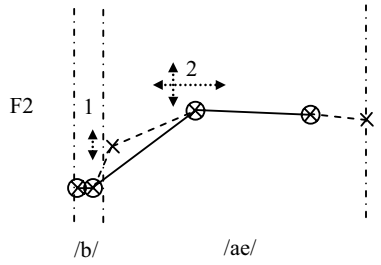

Figure 5. Representation of transition between a voiced labial plosive and a vowel.

Although the formants are not clearly detectable in plosive bursts, they have an important role determining the burst frequency and also the locus point for formant transitions of the following vowel. The short duration of the plosive burst of /b/ does not allow much variation in control points. If the duration is one or two frames the control points of 4-point and 2-point models are exactly the same. The only differentiating point is the value of the first vowel frame. However, the variation of that point (marked with a number '1') can not be too large without introducing an audible discontinuity. The second point of the vowel (marked with a number '2') has a greater effect on the transition speed, which seems to be more important than the details of the transition shape. The context dependent rules can successfully adjust the location of the control points in such a way that the essential and perceptually relevant information is preserved.

6. Conclusions

The past research has shown that the exact modeling of the formant contours is not necessary for correct vowel identification in synthetic speech. Our study investigated the possibility of reducing the complexity of formant contour modeling in formant synthesis. The baseline implementation used four control points per phoneme and context dependent rules in assigning their values. A listening evaluation test was carried out to compare the 4-point model against the 2-point model where the first and the last control points were omitted. The test included 23 utterances synthesized in US English. The results showed that there was no perceptual difference in synthetic speech quality between the two versions. The results are explained partly by the psychoacoustic effects of human auditory system and also by the modeling capabilities of the dynamically assigned control points. The two control points which can be positioned in different places along the duration of the phoneme are able to model rapid transitions between phonemes (onset and offset) as well as the formant shifts in diphthongs. The use of the 2-point model reduced the number of slope calculations in linear interpolation by two and provided 13% memory reduction of language-specific rules.

7. References

[1] Klatt, D. H., "Software for a cascade/parallel formant synthesizer", *J. Acoust. Soc. Amer*., *67(3), 1980, p 971.*

[2] Klatt, D. H. and Klatt, L. C. "Analysis, synthesis, and perception of voice quality variations among female and male talkers", *J. Acoust. Soc. Amer*., *87(2), 1990, p 820.*

[3] Iso-Sipilä, J., Viikki, O., and Moberg, M., "Multi-Lingual Speaker-Independent Voice User Interface for Mobile Devices", *In Proceedings of ICASSP 2006*, Toulouse, 2006.

[4] Allen, J., Hunnicutt, M. S., and Klatt, D., "*From text to speech, The MITalk system*". Cambridge University Press, Cambridge, 1987.

[5] Hertz, S. R., "The delta programming language: An integrated approach to non-linear phonology, phonetics and speech synthesis", *Papers in Laboratory Phonology I,* Cambridge University Press, 1990.

[6] Rabiner, L. R., "A Model for Synthesizing Speech by Rule", *IEEE Transactions on Audio and Electroacoustics*, Vol. 17, Issue 1, pp. 7-13, 1969.

[7] Hillenbrand, J. M. and Nearey, T. M., "Identification of resynthesized /hVd/ utterances: Effects of formant contour", *J. Acoust. Soc. Amer.* Vol. 105, Issue 6, pp. 3509-3523, 1999.

[8] Neel, A. T., "Formant detail needed for vowel identification", *Acoustics Research Letters Online (ARLO)*, Vol. 5, Issue 4, pp. 125-131, 2004.

[9] Pärssinen, K. and Moberg, M., "Multilingual Data Configurable Text-to-Speech System for Embedded Devices", In *Proceedings of Multiling 2006*, Stellenbosch, South Africa, 2006.

[10] Breiman, L. et al., *Classification and Regression Trees*. Belmont, CA, USA: Wadsworth Inc., 1984.

[11] Taylor, P., Black, A. and Caley, R., "The Architecture of the Festival Speech Synthesis System", In *Proceedings of the 3rd ESCA Workshop on Speech Synthesis, Jenolan Caves, Australia.* p. 147-151, 1998.

[12] Stevens, K. N., "*Acoustic Phonetics*", Cambridge, Massachusetts, The MIT Press, 1998.

[13] Ball, M. J.; Rahilly J., *Phonetics, the science of speech*. New York, USA: Oxford University Press Inc., 1999.

[14] Kincaid, D. and Cheney, W., "*Numerical Analysis, 2nd Edition*", Brooks/Cole Publishing Company, 1996.