WORD MODEL-DETERMINED SEGMENTAL DURATION IN FINNISH SPEECH SYNTHESIS AND ITS EFFECT ON NATURALNESS

Jussi Hakokari[1], **Tuomo Saarni**[2], **Mikko Jalonen**[2], **Olli Aaltonen**[1], **Jouni Isoaho**[2], **Tapio Salakoski**[2]

[1]Phonetics Laboratory, Department of Finnish and General Linguistics, University of Turku ([Turku]Finland)

[2]Department of Information Technology, University of Turku ([Turku]Finland)

Abstract

The 50-year-old concept of formant synthesis was under much scientific scrutiny until it became apparent that naturalness was hard to attain using rule-based synthesis methods. Formant synthesis was essentially an intelligible approximation, rather than imitation, of human speech. Our current research is a revisitation to the rule-based formant synthesis. At the moment, due to advances in computer technology, we are in a better disposition to develop naturalness in rule-based text-to-speech systems. While research teams in various countries have made similar efforts, our project is unique in Finland and, more importantly, as regarding the Finnish language. The duration of individual phones is important to naturalness in Finnish. We have extracted data from an extensive single-speaker Finnish speech corpus and created word models to prescribe each phone a duration depending on its position within a word. In this paper, we will describe our approach, present the preliminary results of listening tests and discuss the potential of word models in improving naturalness in text-to-speech systems.

Keywords: rule-based, formant, text-to-speech, consonant/vowel, pattern, Klatt, TTS

1. Introduction

The Finnish language exhibits contrast between phonemically short and long segments (also called chronemic contrast). This contrast applies to all vowels and most consonants. The short vowels are generally more central in vowel space while the long ones are peripheral (Wiik 1965, Lennes 2003). The decisive factor, however, is duration (Wiik 1965) and Finnish speakers are unaware of any qualitative differences between the two chronemic variants. The acoustic difference between short and long phonemes is not linear, but relative to the segment's position in the syllabic structure of the word, and to some degree, the word's position within a sentence. The aim of the listening test in this study is to investigate whether or not varying segmental duration is useful in improving naturalness and rhythm in a TTS (text-to-speech) application. Our long term objective is to prepare a TTS system that will not only introduce varying mean durations, but also quantitative and qualitative reduction characteristic of natural speech as well as sentence context sensitive modeling of duration and fundamental frequency.

2. Methods

2.1. Data analysis

We have examined data on segmental durations presented in Lehtonen (1970), and datamined a single speaker speech corpus of 692 segmented and annotated sentences prepared and studied by Vainio (2001). The corpus contains approximately 6500 words, and is read aloud by a 39-year-old male, a native Finnish speaker from Helsinki. Sentence lengths in the corpus vary from 2.18 s to 20.00 s, and the database adds up to approximately 69 minutes of recording.

All the consonant/vowel patterns of individual words were extracted from the speech corpus automatically using software designed specifically for the task. The software makes use of the original annotation provided with the corpus, and allocates each consonant/vowel pattern found indiscriminately into its own class, maintaining duration information of each phone. A consonant/vowel pattern, together with the mean durations of each segment is referred to as 'word model'. For instance, our data contained 125 occurrences of VCCV – pattern words such as <usko> (*faith*) and <akka> (*an old woman*). Any word with a single short vowel, a geminate or two consecutive consonants, and finally another single short vowel will fall into this category. Our data added up to a mean duration structure of 78 ms for the first vowel, 61 ms for the first consonant, 66 ms for the second consonant (or a total of 127 ms for a geminate), and 48 ms for the final vowel. We have been able to establish ~1100 different word models.

To implement, we have prepared a synthesizer that automatically determines each segment's duration by matching the word against its corresponding model in the database. For instance, in CVCCV words, such as <miksi> (*why*), the first vowel has a mean duration of 73 ms, whereas the second has a mean of only 53 ms (208 tokens). The synthesizer produces the closest match possible to the values in the database; the duration of an individual segment varies to some degree due to F0-induced differences in wave length. We are unaware of any other Finnish TTS taking that kind of within-word environment into consideration.

2.2. Stimulus generation

The first set of stimuli for the listening test was produced using the original, unaltered configuration of the synthesizer which produces speech signal with fixed segmental duration. There is a cascading F0 contour (100-120-80 Hz). The second set of of stimuli was generated with an improved model, that introduces more variation in F0 (100-140-80 Hz) and word modeling to determine segmental durations. The original configuration produces greater segmental and overall durations (30–35 % longer) than the improved one; the first set of stimuli was adjusted to the same length with the second using a PSOLA (Pitch-Synchronous Overlap and Add) algorithm. The operation maintains the spectral characteristics and fundamental frequency of the signal. Additionally, we have had to make minor adjustments to how transitions between phones are realized because the word models cannot be implemented to original system as such. The 16 stimuli represented four categories: four single words (0.74 s – 1.03 s in duration), four short sentences containing short words (1.11 s – 2.45 s), four longer sentences with long words (3.64 s – 4.00 s), and four sentences of medium length with no particular constraints (2.38 s – 2.71 s). All of the sentences were in Standard Finnish and adapted from newspaper articles; some of the sentences represented informal literary style while the majority were formal.

All the synthetic stimuli were produced with the data driven formant synthesis program under development at the University of Turku. The program uses SenSyn 1.1

software for signal generation from a parameter file. SenSyn 1.1, by Sensimetrics Corporation, is based on KLSYN88 synthesizer (Klatt 1982). The resulting signal has a sample rate of 10 kHz. The organization structure of the overall system is illustrated in figure 1.

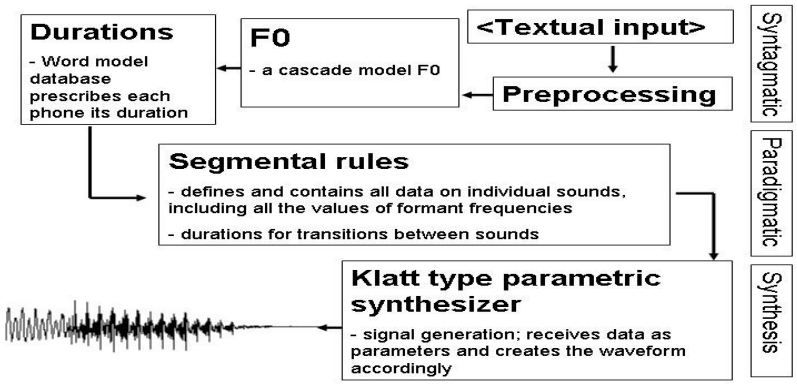

Figure 1. A modular representation of the TTS system

2.3. Participants

None of the participants had been exposed to the synthesis in question previously, and they were neither specialists in language nor synthetic speech. They were asked about their primary and secondary dialect background, since prosody, segmental duration included, is dialect sensitive in Finnish. There were 10 women from ages 20 to 54. Two of the participants were left-handed, and no one reported any deficit in hearing or language.

2.4. Listening test procedure

Due to similarity and brevity of the stimuli and naïve participants we opted for a forced choice paradigm instead of Category Estimation as the evaluation method. The participants heard two words or sentences of identical length successively. Their task was to identify which one of the two sounded more natural. They had transcripts of the sentences to prevent intelligibility issues from diverting them from their task. They were specifically instructed to judge how well the stimuli corresponded to human speech patterns, instead of how clear, pleasant, or intelligible they were. The participants were presented the stimuli in a pseudorandomized order, so that the original and improved versions would not occur consecutively. Presentation order was the same for all participants. The participants judged a total of 16 stimulus pairs.

The session lasted for 15 minutes, and took place in an ordinary laboratory room with no external distraction or noise. The uncompressed sound files (.wav) were played with an ordinary laptop computer and Labtec LCS-1060 loudspeakers. Volume was adjusted to be as loud as possible without causing distortion in the signal or discomfort in the participants.

3. Results and discussion

By data analysis, we have found that even in perfectly intelligible sentences read clearly aloud there is considerable overlap between short and long phonemes. In other words, a

short phoneme may be longer in duration in one position than a long phoneme in another, and that alone does not cause intelligibility issues. For instance, the short vowel /o/ varies from a (reduced) single periodic waveform to 200 ms, while the long vowel /o:/ varies from 54 ms to 294 ms. The overlap applies to all the phonemes in the corpus which exhibit chronemic contrast. Since perception of natural speech is adapted to relative segmental duration, instead of relying on absolute duration, we expected word models to affect naturalness in synthetic speech as well.

The results of the listening test, presented in figure 2, were generally ambivalent towards the use of word models. Only 66 of the 160 responses (41,25 %) preferred the improved configuration. 76 (47,50 %) of the responses preferred the stimulus presented first, and 84 (52,50 %) preferred the second; there was no bias concerning order of presentation. 4 out of the 16 improved stimuli were judged better than the original ones, two of them unanimously. 3 were at chance level, and 9 were deemed worse.

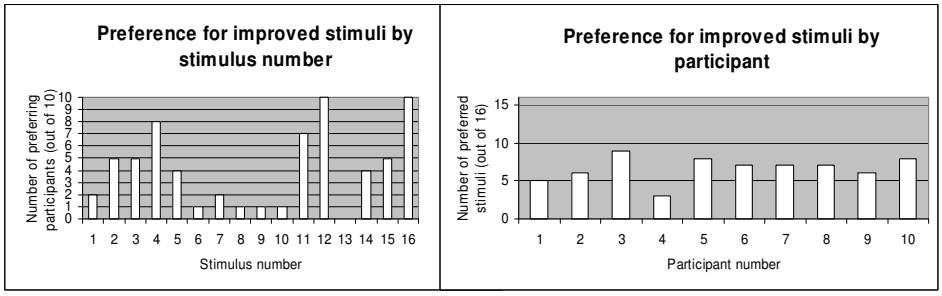

Figure 2. Numerical data

The five least favoured improved stimuli (zero or one preferring responses) were either in the single word or the short sentences with short words category. The two improved stimuli preferred unanimously were of the long sentences with long words category. The other two rated above chance level (eight and seven preferring responses) were of the unconstrained and the short sentences categories. The effect of dialect background is hard to determine due to the homogeneity of the participants. However, the improved set of stimuli was clearly rated lowest by the Southeastern Finnish (South Karelian) speaker, the only non-Southwesterner in the group, who preferred only 3 out of 16 improved stimuli. The following examples, from 1 to 4, are the sentences that were judged more natural in the improved version; the English translations are in italics. Examples from 5 to 9 are the sentences that were judged unnatural the most.

(1) Radion faktasarjan palkinnon sai kulttuuriohjelmien dokumentti. *A documentary by cultural programs (a department of the Finnish Broadcasting Company) won the factual series of the radio competition.*
(2) Hannover on Euroopan tärkeimpiä kansainvälisiä keskuksia. *Hanover is one of the most important international hubs in Europe.*
(3) USA:n (<uuesaan>) joukot toimivat ilman YK:n (<yykoon>) lupaa. *The US forces operate without UN approval.*
(4) Ei liian erikoinen eikä liian tavallinen. *Not too special or ordinary.*
(5) Sunnuntaisin. *On Sundays.*
(6) Peruskorjaus. *Renovation.*
(7) Osakkeenomistaja. *Shareholder.*
(8) Hän ei ole enää olemassa. *(S)he exists no more.*

(9) Miksi Turku ei kasva? *Why isn't Turku getting bigger?*

In the light of present data, word modeling does not appear to improve naturalness universally; the results show responses below chance level. However, we can see improvement in the category of long sentences with long words. We can identify several possible causes for the conflicting results.

At this point it is unclear how the word models will affect once other naturalness features, prosodic and segmental, are implemented. The current system incorporates word models into a synthesizer that is designed to handle fixed segmental durations. Less peripheral formant values typical of ordinary speech (Lennes 2003), a better modeling of F0 contours, and somewhat longer segmental durations might suit the word modeled synthesis better. At the moment the synthesis uses highly peripheral (great acoustic distance between speech sounds) formant values to promote intelligibility.

The more common models, essentially short words, are based on mean values calculated from a sample size up to 287 tokens (the model CV). The longer and strongly inflected word forms, the word models of which are based on only one token (an occurrence of the word in the corpus), make up ~59 % of the database. Those word models carry greater within variation in segmental duration, while within variation has been neutralized due to averaging in the others. That may explain why longer words scored better in the preliminary experiment. In future experiments, it would be worth the while to base all word models on single tokens. The current database is primitive in that it treats all consonants and all vowels equally. In addition, there is no distinction of sentence environment; words occurring in the beginning and the end of sentences are all included into the database without any special tagging. A more detailed datamining could produce contextual classes for the word model database that would differentiate, for instance, voiceless stops from fricatives.

Stimuli used appeared to have too fast an articulation rate (up to 407 syllables per minute). The participants reported they had difficulties in judging the stimuli. They may have found fast, synthetic speech they are unaccustomed to confusing, and picked up a strategy that favors one set of stimuli over the other by some factor other than rhythm and naturalness of speech. The second set of stimuli could be lengthened to match the overall duration of the first in a future study, instead of shortening the first set. This would make the task of judging synthetic speech easier to the naïve participant. Another line of study could make use of prolonged exposure. First, the participants could judge entire paragraphs of text or newspaper articles instead of single words and sentences. Second, the participants could become familiarized to synthetic speech beforehand to avert confusion. With synthetic speech, certain monotonous or reoccurring elements may become irritating after a while. Word models may help to create a less predictable and monotonous synthesis.

Segmental durations are subject to considerable dialectal variation in Finnish. Even if the word models do not contribute a great deal to speech quality or naturalness per se, it is worth further attention to investigate whether they can be used to emulate speakers from different regions or localities. It may be of consequence from the vantage point of Finnish TTS product development, as some end users may prefer to have the synthesizer speak in a manner familiar to them. If dialectal variation, individual speakers or styles can be imitated by the method, it is, from a scientific point of view, a discovery itself. Nine out of ten participants in the present study reported to speak one of the Southwestern dialects (the remaining one a Southeastern speaker), and none were speakers of Helsinki dialect the word models are based on.

4. Summary

In this paper, we have studied the use of word models to prescribe segmental duration in Finnish language speech synthesis. First, we established an initial set of word models by datamining a single speaker speech corpus. Second, we have implemented the word models into our developing rule-based TTS system. Third, we have done a preliminary listening test to examine their effect on naturalness in synthetic speech.

Our studies show there is indisputable overlap between long and short phonemes. Chronemic contrast has been handled in speech synthesis by giving short and long speech sounds fixed durations. Naturalness may be improved by additional modeling of segmental duration according to how it is realized in natural speech. The listening test, however, showed that the method does not give a straightforward advantage. Instead, the results are ambiguous but suggest the word models are most effective when applied to long words in long sentences. The synthesizer in its current, experimental configuration does not improve naturalness in shorter utterances. There is potential in at least a selective implementation of varying segmental duration, and the matter requires a more detailed and more comprehensive investigation. Word modeling may be sensitive to qualities in synthetic speech we have not yet taken into consideration.

References

Klatt, Dennis 1980. Software for a cascade/parallel formant synthesizer. *Journal of the Acoustical Society of America* 67. 971–995

Lehtonen, Jaakko 1970. Aspects of quantity in standard Finnish. Jyväskylä: University of Jyväskylä.

Lennes, Mietta 2003. On the expected variability of vowel quality in Finnish informal dialogue. In: Solé, M., Recasens, D., Romero, J., (eds.) *Proceedings of the 15th International Congress of Phonetic Sciences (ICPhS).* 2985–2988.

Vainio, Martti 2001. Artificial neural networks based prosody models for finnish text-to-speech synthesis. Helsinki: University of Helsinki.

Wiik, Kalevi 1965. Finnish and English Vowels. Turku: University of Turku.

JUSSI HAKOKARI is a research assistant at the Phonetics Lab (Department of Finnish and General Linguistics) at the University of Turku in Turku, Finland. He received his B.A. (phonetics) at the University of Turku, dealing with speech synthesis. His research interests concern speech synthesis and speech acoustics. His master's thesis focuses on Finnish language TTSs and rule-based formant synthesis. As a lecturer, he has taught acoustic analysis, pronunciation and phonetic transcription at the University of Turku. E-mail: jussi.hakokari(à)utu.fi.

TUOMO SAARNI is a research assistant at the Department of Information Technology at the University of Turku in Turku, Finland. He received his B.Sc. (computer science) at the University of Turku, dealing with visibility algorithms in 3D computer graphics. His research interests concern automatic analysis methods in developing speech synthesis. His master's thesis focuses on Finnish language TTSs, rule-based formant synthesis and data mining of natural speech corpora. E-mail: tuomo.saarni(à)utu.fi