The Role of Duration in Finnish Rule-Based TTS

Tuomo Saarni[1] & Jussi Hakokari[2]
Tapio Salakoski[1], Jouni Isoaho[1] & Olli Aaltonen[2]

[1]Department of Information Technology
[2]Phonetics Laboratory
University of Turku, Turku, Finland

University of Turku
FIN-20014, TURKU
tuomo.saarni@utu.fi & jussi.hakokari@utu.fi

ABSTRACT

We are developing a rule-based Finnish-language TTS system. Our primary concern is to find ways to increase naturalness in the synthesis. Our approach is to observe tendencies in natural language through acoustic analysis and data mining, and to implement our findings into the synthesizer. We have concentrated on modeling duration, which is an essential part of Finnish prosody. The language exhibits contrasting phonemic lengths and the durations of individual phones are highly sensitive to their position within a word. We have developed a duration model ("word models") based on how the syllabic structure of a word correlates with segmental durations in a natural speech corpus. We have implemented and automatized the word models, and studied through listening tests whether they improve naturalness in the synthesis. We compared the word model–determined segmental durations with with fixed ones. The result was ambiguous: the word models appear to improve naturalness in longer speech stimuli, but not in the shorter ones.

1. Introduction

While rule-based speech synthesis is versatile and, when correctly configured, intelligible, its weakness lies in naturalness; it is very difficult to produce synthetic speech that sounds humanlike without resorting to samples of recorded speech and the concatenative methods. Our aim is to investigate just how far one can go with rule-based synthesis by carefully modeling characteristics of natural speech. At the moment we are mostly concerned with duration and its effect on naturalness. Duration is a part of speech prosody and important to a natural rhythm of speech. Duration is also a delicate matter in Finnish speech synthesis; the language is cited as a “quantity language” [3].

The Finnish language has contrasting phonemic length; all the vowels and the majority of consonants may occur either short or long and thus form minimal pairs. The short vowels tend to be slightly more central in the vowel space than the long ones [7, 4], but the decisive factor is duration [7]. Finns are generally unaware of any qualitative differences between the two phonemic lengths. The duration is not absolute even in the widest sense, but relative to the segment’s position within a word and the word’s position within a sentence.

This is a continuation to an earlier study in which we compared ”word model”–determined segmental durations to fixed durations [1]. Word models, inspired by Lehtonen’s work [3], are mean durations based on consonant-vowel sequences data mined from natural speech corpora. For instance, in our training corpus the word form VC (a vowel followed by a consonant) has mean durations of 69 ms (V) and 52 ms (C). The synthesizer retrieves the data from the word model bank and makes the vowel ~70 ms and the consonant ~50 ms long in all words of the form VC. Our word model bank had ~1100 entries. The results were encouraging; long sentences with long words in them were deemed more natural than the sentences synthesized with fixed durations [1]. In this study we made a more complex set of word models; plosives are now distinguished from the rest of the consonants and diphthongs are separated from long vowels. Now the speech corpus yielded ~2500 entries. Furthermore, we used much longer samples of synthetic speech than in the previous study. The speaking rate was also slower; a faster speaking rate used in the first experiment proved confusing to the naïve participants. In this study we investigated whether the improved word models enhance naturalness in synthetic speech.

2. Methods

2.1. Data analysis

The speech corpus from which the word models were extracted consisted of 692 declarative Finnish sentences containing ~6500 words. The corpus, described in more detail in [6], is read aloud by a 39-year-old male from Helsinki and adds up to 69 minutes of recording. The corpus was consistently segmented and annotated at word and phone levels to make data mining possible. The previous word model bank which did not differentiate diphtongs and long vowels or plosives and the rest of the consonants was problematic. Consequently, all the words occurring in the corpus were this time established as sequences of plosives (P), other consonants (C), vowels

(V), and diphthongs (VV). Durations for each segment within a model are included into the bank; multiple occurrences of a single model involve calculating a mean duration for each segment. The number of established word models was ~2500, reflecting the long words typical of a highly inflected language with a small phoneme inventory such as Finnish. In fact, the ~2500 word models are not nearly enough for synthesizing free input. For TTS purposes, the user can choose either fixed durations or a generic word model to determine the segmental durations in case the system encounters a word that falls outside the bank. Both the wavelength and the 10 ms time resolution of the signal generator dictate that the durations prescribed by the word models cannot be reproduced with utmost accuracy; the system uses rounding and waveform interpolation in producing a continuous speech signal.

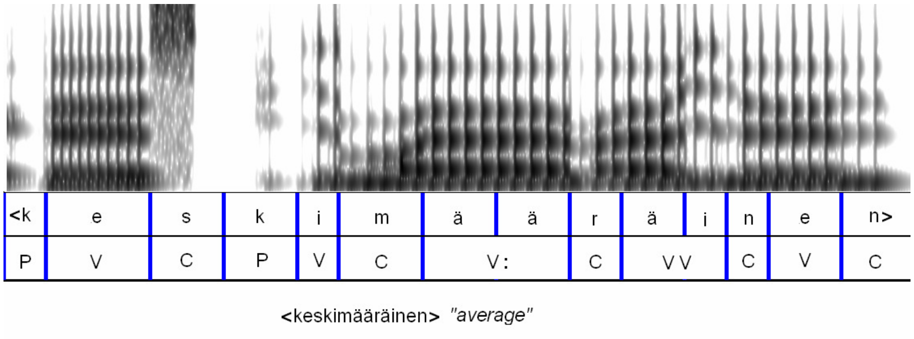
2.2. Stimuli

The eight stimuli were four paragraphs of text synthesized into speech in two different ways. We made sure that all the words in the stimuli had a representation in the word model bank. In other words, the generic models or fixed durations were not used. The stimuli were all Standard Finnish, but one of them (stimulus A) contained two foreign proper names (“Vladimir” and “Visentini”). The first set of stimuli used the word model bank to determine segmental durations. The second set of stimuli used fixed durations based on mean values found in the same speech corpus. The speech rate (= overall duration) of the stimuli was thus practically equal. All the phonemically short segments were ~70 ms in duration, while the long ones were ~140 ms.

Table 1. *Stimulus information.*

	Words	Characters	Duration fixed	Duration word-model
Stimulus A	85	606	44.24 s	42.61 s
Stimulus B	40	261	19.20 s	19.14 s
Stimulus C	75	464	34.13 s	32.62 s
Stimulus D	65	426	31.59 s	31.57 s

The stimuli were sound files (.wav) with a sample rate of 10 kHz. The signal generator is still under development and produces occasional disturbances (pops and clicks) similar to those produced by the Klatt synthesizers [2]. The disturbances were not manually edited out of the signal, but the participants were asked not to pay attention to them.

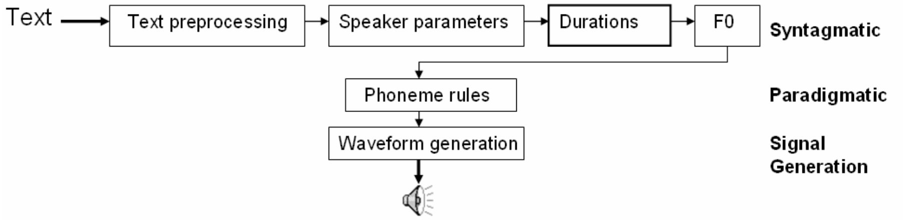


Picture 1. *An example of word model –determined segmental durations in a synthesized word.*

A simple cascading model for fundamental frequency was used in generating the stimuli to highlight the effect of segmental durations. F0 starts at 100 Hz in the beginning of a sentence and climbs up to 140 Hz by the end of the first syllable. Gradually, F0 falls down to 65 Hz by the end of sentence, going up 40 Hz intermittently at each word boundary. Consequently, a drawn F0 contour looks like a gently sloping saw tooth pattern that is tilted towards the left hand side. F0 does not go all the way down to 65 Hz at a phrase boundary within a sentence (i.e. a comma or a semicolon in the input text), but rises 5 Hz in addition to the ordinary 40 Hz rise at word boundaries. There was a 150 ms silence interval at phrase boundaries, and a 350 ms interval at sentence boundaries.

There was also a 46 s test file the participants heard before they began. It was generated using the older word models presented in [1]. There was also a prepausal lengthening module switched on rendering all phrase- and sentence-final words 10 % longer than the rest. Otherwise the configuration was identical to that of the actual test stimuli.

The rule-based synthesizer used for stimulus generation consists of three levels of processing. The syntagmatic level contains preprocessing and a number of modules for prosody and speaker parameters to choose from. The paradigmatic level holds the phoneme and allophone inventories. The third level is signal generation, now handled by JPSyn, a Klatt –type software of our own design.



Picture 2. *The structure of the TTS system.*

2.3. Participants

There were 21 participants, 7 women and 14 men. One of the participants was left-handed. Their average age was 28 years, the eldest being 45 and the youngest 23. The participants were asked about their primary and secondary dialect background, since there is considerable dialectal variation in Finnish speech prosody, segmental durations included. The majority of the participants were speakers of the South-Western dialects of Finnish. One of the South-Westerners was bilingual in Swedish and Finnish. There were six primarily Southern (includes the capital city Helsinki) speakers; an additional three listed a secondary background in Southern dialects.

The participants had to evaluate their experience with synthetic speech in general. The scale extended from 1 (hears synthetic speech several times a week) to 5 (has never been exposed to synthetic speech); the average for the group was 3.5, roughly corresponding to a few times a month.

2.4. Listening procedure

We preferred the test to be taken in a comfortable environment and through a likely medium for speech synthesis use. The participants got to access the entire test material over the internet and carry out the evaluation in their homes. They were instructed to exclude any disturbing noise or movement from their vicinities before beginning and to set the volume in their loudspeakers to a loud but comfortable level. There was also a synthetic 46 s test file (a greeting of a sort) the participants heard first; the test file utilized neither of the duration models under examination.

The test itself was a forced choice paradigm. The participants were instructed to listen to each of the stimulus pairs over as many times as they wanted. They were asked to mark which one of the sentences (A1 or A2, B1 or B2, etc.) they thought sounded more natural and better corresponded to human speech rhythm; the order of the stimuli in a stimulus pair was scrambled, and nothing about the alternative duration models was disclosed to the participants. They were specifically instructed not to let intelligibility issues affect their judgment.

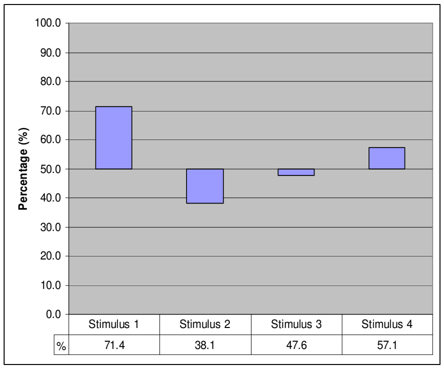
Finally, they were asked to submit their personal information, including age, sex and handedness, and to identify their primary and secondary dialect backgrounds. They would also estimate their amount of personal experience with synthetic speech. The participants submitted their results and information by e-mail using an electronic answer sheet.

3. Results

The results show that the participants preferred the stimuli with word modeled segmental durations only slightly (53.6 % of the stimuli). The first and the longest one of the stimuli was preferred the most (71.4 % for the word models). The second and shortest one of the stimuli was the least preferred (38.1 % for the word models).

Men preferred the word models more than women (58.9 % vs. 42.8% of the stimuli). Four of the participants preferred all the instances of the word models, while two of them preferred all of the fixed duration stimuli; no common denominer was found for them. Dialect background had no significant effect, but those more experienced with synthetic speech (reported an experience level of 2 or 3) were likely to prefer fixed durations (40 % for the word models). The least experienced ones (experience level of 4 or 5) were likely to prefer the word models (65.9 %); they rated the first stimulus better 90.9 % of the time. Three of the participants had preferred the stimulus they heard last in all four cases; this may or may not represent an unconscious bias, but it does not affect the outcome significantly.

Table 2. *Preference for the word-modeled stimuli and partial raw data (1=preference for the word models).*



	Stimulus 1	Stimulus 2	Stimulus 3	Stimulus 4
Participant 1	1	1	1	1
Participant 2	0	1	0	0
Participant 3	1	0	0	1
Participant 4	1	1	0	1
Participant 5	0	0	0	1
Participant 6	1	0	1	0
Participant 7	1	0	1	0
Participant 8	1	1	1	0
Participant 9	1	1	1	1
Participant 10	1	0	0	1
Participant 11	1	1	0	1
Participant 12	1	1	1	1
Participant 13	0	0	0	1
Participant 14	1	0	1	0
Participant 15	1	0	1	1
Participant 16	1	1	1	1
Participant 17	1	0	0	1
Participant 18	0	0	1	0
Participant 19	1	0	0	0
Participant 20	0	0	0	0
Participant 21	0	0	0	0

4. Discussion

The results show that the word models either enhance naturalness (stimulus A), hinder it (stimulus B), or have no effect at all (stimuli C and D). It appears that the longer the synthesized sample is, the more the word models enhace naturalness. That is in line with the preliminary findings in the previous study [1]. Several weaknesses can be identified in the word model approach. First, the word models require a large database. The ~6500 words in the corpus produced ~2500 models. The syllabic structure of Finnish, a highly inflected language, is so complex, that establishing an adequate database would require a much greater corpus. A complete set of word models would require an astronomical amount of entries, since there is no theoretical upper limit for the length of word forms in written Finnish. It may be of interest to examine how the word models perform in a language which requires a limited set of word models (shorter words and less inflection).

Second, the fact that rule-based synthesis is computatively non-expensive and requires little memory capacity is one of the greatest advantages of the method. To implement a vast database would be a compromise in the latter respect. Third, word models represent a single speaker's speaking style. That may be seen as a disadvantage, if one wants to create a generalized, impersonal speaker. On the other hand, a TTS system might contain several speaker profiles with corresponding individual or dialect-specific segmental durations. Fourth, word models based on a large sample size tend to lose some of their shape due to averaging. Conversely, a word model that is based on a single token may reproduce effects of syntactic environment or information structure that are ill-fitted to other contexts.

The overall preference for the word model–determined durations is so weak, that their implementation is not necessarily justified considering the weaknesses. Fixed durations appear to do well in comparison even though they are counterintuitive; it is unlikely a natural language would operate with fixed durations. In fact, there is a chance that duration is not that important from the vantage point of speech perception. People are generally unaware that the acoustic correlate of phonemic length, duration, is relative. They are surprised to hear that within just one word, a phonemically long vowel may be shorter in duration than another short vowel. During speech perception, the brain apparently registers each speech sound as either long or short. The only thing that catches one's ear in the synthesis is when a segment is abnormally short or long; that happens occasionally with the current word models. Carefully modeled phonemes, transitions, F0, and the oft-neglected intensity may prove to contribute more to naturalness than duration. Sakamoto and Saito [5] studied synthetic speech modeled after donor speakers (VoiceFonts), and found that duration has a relatively small effect on speaker recognizability compared to other variables.

We have developed another model that combines the phonemes' intrinsic durations (some are longer than others on the average) with a generic word model. The generic word model makes the syllables grow shorter towards the end of the word, a tendency observed in the corpus as well as in Lehtonen's material [3]. In addition, there is a prepausal lengthening effect. If avoiding large databases is no question of principle, one could of course bring in more variables and create a very extensive word model bank. For one thing, the word models could be sensitive to syntactic roles (necessitates a syntactic parser for data mining purposes). Alternatively, the word model could cover only a limited sequence of phones, for instance the first eight in any word. The remaining phones would be dealt with using a generic model. Nevertheless, we are discouraged to continue the current line of investigation into word models. We are inclined to find alternative methods to model segmental durations in rule-based text-to-speech synthesis.

5. Conclusion

We compared fixed segmental durations to those measured from a natural speech corpus. The results show that word modeled durations improved naturalness only partially according to the listeners' judgment. Therefore, we suggest that the syntactic structure of the sentences should be taken into consideration if the word model approach was developed further. The statistical analysis of duration in various syllabic structures alone is inadequate at least for a language such as Finnish.

References

[1] Hakokari, J., Saarni, T., Jalonen, M., Aaltonen, O., Isoaho, J. & Salakoski, T., 2005. Word-model determined segmental duration in Finnish speech synthesis and its effect on naturalness. M. Langemets & P. Penjam (Eds.*) Proceedings of the 2nd International Conference on Human Language Technologies.* Tallinn: Raamatutrükikoda. 137-142. Available online at http://users.utu.fi/tuiisa/pubs/

[2] Klatt, D., 1980. Software for a cascade/parallel formant synthesizer. *Journal of the Acoustical Society of America 67*. 971-995.

[3] Lehtonen, J., 1970. *Aspects of quantity in standard Finnish*. University of Jyväskylä.

[4] Lennes, M., 2003. On the expected variability of vowel quality in Finnish informal dialogue. M. Sóle, D. Recasens & J. Romero (Eds.) *Proceedings of the 15th International Congress of Phonetic Sciences (ICPhS)*, pp. 2985-2988.

[5] Sakamoto, M., Saito, T. 2002. Speaker recognition evaluation of a VoiceFont-based text-to-speech system. *Proceedings of 7th International Conference on Spoken Language Processing,* pp. 2529-2532.

[6] Vainio, M., 2001. *Artificial neural network based prosody models for Finnish text-to-speech synthesis*. Helsinki: Yliopistopaino.

[7] Wiik, K., 1965. *Finnish and English vowels*. Turku: University of Turku.