

Building a Finnish Unit Selection TTS system

Hanna Silen, Elina Helander,
Konsta Koppinen, Moncef Gabbouj

Institute of Signal Processing
Tampere University of Technology, Finland

{hanna.silen|elina.helander|konsta.koppinen|moncef.gabbouj}@tut.fi

Abstract

Speech synthesis based on unit selection can produce far more natural speech than conventional diphone-based methods. Unit selection based text-to-speech synthesizers have been built for many different languages. In this paper, we describe the development of TUT_VOICE, the first Finnish unit selection synthesis engine for academic research. The system includes database construction, synthesis engine implementation and optimization for Finnish.

1. Introduction

Unit selection [10] is a method of corpus-based concatenative speech synthesis. It uses a large pre-recorded speech inventory to provide a sufficient phonetic and prosodic coverage for a language. Speech is produced by cutting and concatenating units from the database.

One of the major challenges is how to select units from the database. The selection process is guided by two costs, a *target* and a *join cost* [10]. The target cost estimates similarity between a candidate unit and a desired unit and join cost measures the concatenation quality of two consecutive units in terms of the continuity of the spectrum, F_0 and energy.

When appropriate units are chosen in synthesis, contexts are taken into account. However, the quality of synthesized speech depends highly on the coverage of the database. One basic idea of unit selection is to avoid signal processing modifications i.e. prosodic modifications. This poses a challenge to the inventory; it must provide not only a complete coverage of synthesis units but many instances of a same unit in different contexts. The design of the database is thus important and should be tailored to language-specific requirements. In addition, the style of the synthesized speech follows the style of the database.

There has been a vast amount of research on unit selection TTS (text-to-speech) and voices have been developed for many languages. Although there exist a fair amount of freely available research and speech analysis tools (e.g. Festival [3]), for a new language a proper database is still needed as well as rules for grapheme-to-phoneme conversion and linguistic parsing etc. Previously, a Finnish diphone voice (hy_fi_mv_diphone [8]) has been built in Festival [3]. However, no prior (academic) research has been devoted to building a unit selection voice for Finnish. In addition, no suitable database has been available.

This paper describes the process of building a Finnish prototype open-domain unit selection system called TUT_VOICE. The building process of TUT_VOICE consisted of two phases. The first part, inventory construction, involved prompt selection, recording of the inventory, utterance labeling and fea-

ture extraction. In the second part, a unit selection synthesis engine was implemented consisting of target construction, unit sequence selection and waveform concatenation. For TUT_VOICE, some ideas from Festival were adopted but the system was built to work independently from it.

2. Finnish phonetics and phonology

Although a unit selection synthesizer can be built on having little or no knowledge of the language (e.g. [13]), understanding the characteristics of the language is important for good quality TTS. Some issues presented in this chapter helped us to understand some "errors" occurring in the synthesized speech. This chapter outlines the basic principles of Finnish phonetics and phonology from the viewpoint of what is needed for building a speech synthesizer.

2.1. Phoneme system and orthography

Phonemes are typically divided into consonants and vowels. There are eight vowels in Finnish: /a/, /e/, /i/, /o/, /u/, /y/, /ä/, and /ø/. Vowels can occur both short and long and form sequences and diphthongs. Compared to other languages, the number of vowels in Finnish is high [7]. On the other hand, a relatively low number of consonants exists. The low number of consonants enables the appearance of a high number of allophones [7]. The consonants and allophones are summarized in Table 1. Consonants are marked using the notation of Finno-Ugric transcription instead of the International Phonetic Alphabet (IPA). Most of the consonants in Table 1 can form geminates. In addition, the consonants /b/ /g/ /f/ /ʃ/ occur only in relatively new loanwords.

Finnish orthography is phonemic: each phoneme corresponds to a certain grapheme and allophones are not pointed out. Short phoneme quantities are written with a single grapheme (e.g. *i*) whereas long phoneme quantities (e.g. *ii*) and diphthongs (e.g. *au*) with two graphemes. There is only one exception: the orthographic correspondent for the phoneme /ŋ/ is *ng*. The main differences between the Finnish orthography and pronunciation are due to assimilation and boundary gemination [7].

2.2. Syllables and syllabification

Every word can be divided into one or more syllables. Each syllable in Finnish has a vowel as a sonant, i.e. every syllable must contain at least one vowel. The syllable structure list is given in Table 2. Letters C and V denote consonant and vowel, respectively. The notation VV denotes a long vowel or a diphthong. The structure of the most common Finnish syllables is simple and no complex consonant clusters exist as Table 2 shows. The

Table 1: Finnish consonants and their allophones.

	phoneme	allophones	examples
plosive	p	[p]	<i>pallo</i>
	t	[t] [t̚]	<i>tutti, tutit</i>
	k	[k] [k̚]	<i>katu, kirje</i>
	d	[d]	<i>lyhde</i>
nasal	m	[m] [m̥]	<i>maila, kamferi</i>
	n	[n] [n̥]	<i>onni, päähänpisto</i>
		[ɱ] [ɱ̥]	<i>fanfaari, tunti</i>
		[ŋ] [ŋ̥]	<i>kenkä</i>
	ŋ	[ŋ] [ŋ̥]	<i>kengät, kangas</i>
fricative	s	[s]	<i>sana</i>
	h	[h] [h̥]	<i>tahto, raha</i>
lateral	l	[l] [l̥]	<i>lika, laki</i>
trill	r	[r] [r̥]	<i>penger, taru</i>
		[r̥]	<i>tutkimusretki</i>
approximant	v	[v]	<i>vanha</i>
	j	[j]	<i>juhla</i>

majority of the words are polysyllabic. Only 17 % of the words in the sentence set used in database construction (Section 3) were monosyllabic. For comparison, the respective number for English calculated from CMU ARCTIC database [2] was 72 %.

Table 2: Finnish syllable types.

common	CV	CVC	CVV	CVVC	VC
	V	VV	CVCC	VVC	VCC
rare	CCV	CCVC	CCVV	CCVVC	CCVCC

The syllabification rules for Finnish are simple and no dictionaries are needed. Only foreign words and compounds words can cause some exceptions. According to [7], Finnish syllabification can be carried out using the following rules:

- A syllable boundary appears before every sequence CV (e.g. *ka-tu*)
- A syllable boundary appears inside every sequence VV unless the sequence is a diphthong or long vowel (e.g. *a-lu-e*)
- A vowel sequence VV ending with /i/ is a diphthong if it is not in the first syllable (e.g. *u-te-li-ai-suus*)
- A vowel sequence VV ending with /u/ or /y/ and not located in the first syllable can be realized as a diphthong or a vowel sequence (e.g. *sel-ke-ys, sel-key-teen*).

2.3. Prosody

Prosody is not expressed through simple phonetic segments but larger units like syllables, words, sentences or even paragraphs. Prosodic features, such as quantity, stress and intonation play an important role in conveying information. It is generally believed that the naturalness of synthesized speech is improved through better prosody modeling. Although our unit selection synthesizer does not explicitly model prosody, there is a need to extract linguistic features that are assumed to affect the synthesized prosody. For example the Finnish diphone voice built for Festival [8] sometimes sounds like a foreigner speaking Finnish due to strange prosody. Thus understanding of Finnish prosody can help to optimize the database and come up with meaningful target costs for synthesis.

One important manifestation of prosody is quantity. It can be determined physically or linguistically [1]. Physical quantity corresponds to a duration of a phoneme while linguistic quantity describes how a native speaker perceives the length. In Finnish there are two distinctive quantities: short and long for both vowels and consonants (geminate). The ratio of short and long vowel durations often differs from 1 : 2 [1].

Finnish word stress is fixed. The primary stress is always on the first syllable while the second and the last syllable are unstressed. In longer words, secondary stress can occur as well.

Voice quality can be also considered as a dimension of prosody. In Finnish, the use of creaky voice at least at the end of a sentence is a frequent phenomenon although it can also appear elsewhere in the sentence [12]. Diphone-based synthesizers avoid the problem of creaky endings, since diphones are extracted from a stable speech section and they are modified to be suitable for every part of a sentence. However, a unit selection synthesizer faces the problem since all speech material is used for synthesis and for example TUT_VOICE does not carry out prosodic modifications.

3. Database and voice construction

The lack of appropriate speech databases is a major problem for smaller languages like Finnish. An important output of this project is a speech database consisting of 1003 utterances optimized for TTS synthesis and spoken by a female speaker. The sentences are narrative and read in a vivid style, since we aimed at expressive prosody. Hence, the database should also should be useful for prosody research purposes.

3.1. Prompt design

The speech inventory was designed for Finnish unit selection synthesis using diphone-sized units. The design followed the idea of the English CMU ARCTIC databases [2]. In total 33 Finnish out-of-copyright books with 203 339 sentences were extracted from Project Gutenberg [4]. Altogether 46 067 sentences with 6-15 words were used as source data for the greedy prompt sentence selection.

Short and long phoneme quantities were treated as different phonemes in the selection. Due to the high allophonic variation of the Finnish consonants, a diphone variant-based approach was taken. By the concept of a diphone variant we distinguish diphones with similar phone content but variation in allophone content or syllabic position. For example, the inter-syllabic diphone *n.k* in the word *vanha* ([van-ha], *old* in English) is considered different from the inter-syllabic variant in the word *vanki* ([vaŋ-ki], *prisoner*) as well as the intra-syllabic variant in the word *vana* ([va-na], *trail*). If the variants are ignored in the prompt selection, there is no guarantee, that all the variants are included in the inventory.

Two separate sets of prompt sentences were selected. The first set (Set A) was optimized to provide full coverage of diphone variants occurring in the source data. New sentences were included as long as there were diphone variants missing from the inventory. Since boundary gemination makes it difficult to predict the actual pronunciation on the word boundaries, inter-word diphone variants were ignored in optimization. Due to vowel harmony in Finnish, front and back vowels do not appear in the same word and inter-word diphone variants consisting of vowels were taken into account.

The second set (Set B) was designed to be rich in syllables. Allophonic variation and stress were taken into account

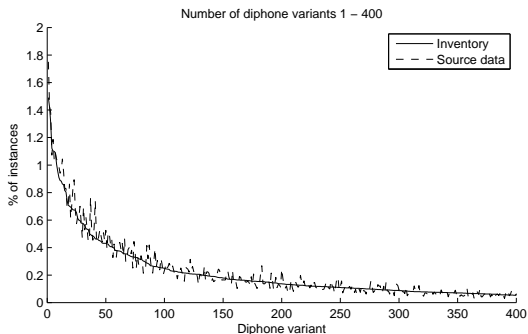


Figure 1: *Distribution of the most frequent diphone variants in the inventory.*

in the selection. Stress in Finnish is fixed and word-initial syllables were considered stressed and the word-final unstressed. Sentences were greedily selected by choosing always the one providing the largest amount of new syllables. The last word of each sentence was ignored due to the possible occurrence creaky endings. The first word of each sentence as well as the monosyllabic words were ignored as well. Sentences of the Set A were taken into account in the selection. After manual removal of archaic and foreign sentences, a set of 1003 prompt sentences was left. Sentences were recorded with a female voice using a sampling frequency of 32 kHz.

The distribution of the most frequent diphone variants in the inventory is illustrated in Figure 1. The solid line denotes the percentage of diphone variant instances of all the instances in the inventory while the dashed line denotes the corresponding value for the source data. As can be seen, the inventory distribution follows well the source data. The 440 most frequent diphone variants cover 90% of the inventory.

3.2. Automatic labeling

Automatic labeling of the inventory utterances used scripts of the Multisyn build tool [9] with slight modifications. HMM-based (hidden Markov model) phoneme models were trained with HTK (Hidden Markov Model toolkit) [15] and forced alignment was used for the phone boundary determination. Boundary alignment was done by using 5-state monophone HMMs. Plosives were divided into closure and explosion phases and separate 4-state HMMs were trained for them. Diphthongs turned out to be very difficult for the alignment and were therefore trained as separate models instead of separating the phones. Diphone boundaries were computed as the mid-point between the phone boundaries except for the plosives which had an aligned boundary between the closure and explosion.

Inventory utterances were spoken relatively fast which complicated automatic labeling. Some of the phones were very short, such as //, /j/, and vowels belonging to diphthongs. They get fused together and even manual labeling of these phones turned out to be difficult. In synthesis, // and /j/ should be extracted as triphones rather than diphones. However, this would require including more data in the inventory in order to guarantee full coverage.

4. Synthesis engine implementation

The TUT_VOICE synthesis engine was implemented as a prototype unit selection TTS system for academic use. The im-

plementation was inspired by the Festival TTS framework [3]. Adding new voices is easy and requires no system compilation. Adjusting the synthesis parameters such as target and join subcost weights as well as changing the grapheme-to-phoneme rule sets can be done without compilation. The core system is implemented for Linux in C++ and the voice construction scripts in Perl. Examples of synthesized speech are available at [14].

4.1. Target construction

Grapheme-to-phoneme mapping for Finnish is quite straightforward and syllabification is done based on some simple rules described in Section 2. The structure of the input sentence is determined by parsing the sentence into a tree-like form similarly to Festival.

4.2. Unit sequence search

Selection of the candidate unit sequence is carried out by computing the total cost $C(\mathbf{t}, \mathbf{u})$ between the target unit sequence \mathbf{t} and a candidate unit sequence \mathbf{u} [10] as

$$C(\mathbf{t}, \mathbf{u}) = \sum_{i=1}^N C^t(t_i, u_i) + \sum_{i=2}^N C^j(u_{i-1}, u_i). \quad (1)$$

Here $C^t(t_i, u_i)$ denotes the target cost between a target unit t_i and a candidate unit u_i and $C^j(u_{i-1}, u_i)$ the join cost between candidate units u_{i-1} and u_i . The best candidate unit sequence \mathbf{u}^* the one that minimizes the total cost, i.e.

$$\mathbf{u}^* = \arg \min_{\mathbf{u}} C(\mathbf{t}, \mathbf{u}). \quad (2)$$

Optimization is done by using the Viterbi search algorithm [10].

4.3. Target cost

The target cost is used to estimate the dissimilarity of a target unit and a candidate unit from the inventory. It is formed as a weighted sum of subcosts. Subcosts are selected in a way that they can characterize phonetic and prosodic properties of the units.

The formula for the target cost $C^t(t_i, u_i)$ of a target unit t_i and a candidate unit u_i is [10]

$$C^t(t_i, u_i) = \sum_{n=1}^q w_n^t C_n^t(t_i, u_i), \quad (3)$$

where q denotes the number of subcosts $C_n^t(t_i, u_i)$ and w_n^t a weight given to a subcost C_n^t .

The used subcosts were the position in syllable, word, and sentence; stress; and left and right phoneme context. All the subcost weights were manually adjusted.

As in Festival, the highest weight was given to stress. Three different cases were distinguished: syllables with primary and secondary stress and syllables with no stress. Since the primary stress is always on the first syllable in Finnish, word stress is related to word boundary detection [1]. Stress was therefore considered important in terms of intelligibility as well.

The second highest weight was given to the unit's position in a sentence. High weight was used in order to avoid the selection of the candidate units with a creaky voice for the target units not from sentence-final words. A unit's position in a syllable and word were not considered as important and were less highly weighted. Units were considered either sentence/word/syllable initial, medial, final, or word/syllable inter.

Table 3: *Synthesis parameters for the target subcosts.*

target subcost	feature values	weight
position in syllable	{initial, medial, final, inter}	0.1
word	{initial, medial, final, inter}	0.1
sentence	{initial, medial, final}	0.3
stress	{primary, secondary, unstressed}	0.35
phone context		
left	{a, a:, b, b: . . . , oe:}	0.1
right	{a, a:, b, b: . . . , oe:}	0.05

Allophonic variation of the phonemes was not taken into account in the transcription. Instead, coarticulation was estimated by the left and right context subcosts.

4.4. Join cost

The join cost is used to estimate the audible mismatches occurring in unit concatenation. Similarly to the target cost, the join cost is formed as a weighted sum of subcosts. Differences in spectral features, F_0 , and power are typically considered in join cost computation [10]. Formula for the join cost $C^j(u_{i-1}, u_i)$ for candidate units u_{i-1} and u_i is [10]

$$C^j(u_{i-1}, u_i) = \sum_{n=1}^p w_n^j C_n^j(u_{i-1}, u_i), \quad (4)$$

p denoting the number of join subcosts and w_n^j the weight given to each subcost.

A continuous pitch contour on the unit boundaries was achieved by using the distance of the units' F_0 as a join subcost. However, since no F_0 is extracted for the unvoiced segments, the F_0 join subcost between two arbitrary selected unvoiced segments equals zero. Therefore the use of F_0 subcost can not guarantee good overall pitch contour. To overcome the problem, we linearly interpolated the values for the unvoiced parts based on the F_0 values of the surrounding voiced parts. Values were normalized to have mean value of 0 and variation of 1.

Spectral mismatches were estimated by the weighted mean-square error (WMSE) of LSFs (Line spectral frequency coefficients). In comparison to MFCCs (Mel-frequency frequency coefficients), LSFs have turned out to estimate better the occurring audible mismatches [6]. The WMSE of two LSF frames \mathbf{f}_1 and \mathbf{f}_2 given as is computed as [11]

$$d(\mathbf{f}_1, \mathbf{f}_2) = \sum_{n=1}^p w_n (f_1(n) - f_2(n))^2, \quad (5)$$

where w_n denotes the weight and $f_1(n)$ and $f_2(n)$ the n th coefficients of the frames \mathbf{f}_1 and \mathbf{f}_2 , respectively. The weight w_n is given as

$$w_n = \max_{i=1,2} \frac{1}{f_i(n) - f_i(n-1)} + \frac{1}{f_i(n) - f_i(n+1)}. \quad (6)$$

The waveform amplitude was controlled by the power join cost. The subcost value was computed as the absolute difference of the power of one pitch period at the concatenation point. Extracted power values were normalized into range of [0, 1].

Difficulties in labeling of some short and poorly detectable phonemes were compensated by introducing a triphone join

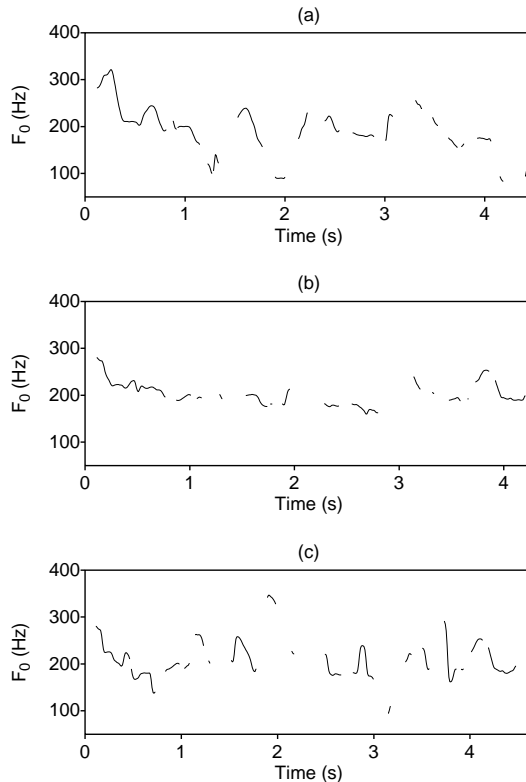


Figure 2: F_0 contours for (a) recorded utterance, synthesised utterance (b) with F_0 interpolation, and (c) without F_0 interpolation.

subcost. Especially the phonemes /l/ and /j/ were found to be very difficult to label correctly, even manually. The aim of the triphone cost was to guide the selection in these cases towards triphone-sized candidate units rather than splitting the short phonemes in order to form diphones. Triphone cost was defined to get a value of 1 if the diphone should be selected as a triphone and 0 otherwise. By using a weighted triphone subcost rather than forced triphone selection, a wider variety of candidate units was achieved.

The weights for each join subcost are listed in Table 4. Differences in weighting indicate the different range of subcost values rather than importance of certain feature.

Table 4: *Weights for the TUT_VOICE join subcosts.*

target subcost	weight
F_0	0.05
LSF	1
power	3
triphone	1

4.5. Pre-selection

In order to speed up the unit sequence search, a pre-selection was used. Units with numerous instances in the inventory were divided into groups of inter- and intra-syllabic instances. Units from the other group were not considered as candidate units and were therefore left out from the search. For the diphones

with less than 5 instances, no pre-selection was carried out. The effect of pre-selection was tested by synthesizing a set of utterances with and without the pre-selection. Among 300 synthesized sentences, 222 were the same regardless of whether the pre-selection was used or not.

4.6. Waveform concatenation

Unit waveforms were extracted from the inventory utterances pitch-synchronously. Diphone boundaries were aligned at the midpoints between the phone boundaries determined by HTK. As an exception, the plosives were divided at the end of the closure determined by HTK. In order to achieve pitch-synchronous waveform extraction, the boundary was moved on the nearest pitch mark.

Glitches on the unit boundaries were avoided by allowing some overlapping. The best concatenation point was determined by finding the width of overlap that provided the highest value of cross correlation. A smooth transition was obtained by averaging the signals in the overlapping region. Roughly one pitch period of overlap was allowed.

Figure 2 illustrates the utterance “*Sehän on jo valmis rautatie, pengertehy, ojat kaivettu, kiskot pantu paikoilleen.*” as (a) recorded, (b) synthesized with F_0 interpolation, and (c) without F_0 interpolation. Note the undesirable jumps in the F_0 in the utterance with no interpolation used in F_0 extraction.

5. Listening experiments

Since the output of a TTS system is speech, the evaluation of a TTS system is usually carried out by conducting listening tests. With speech coders, a MOS test is commonly used and it has been applied to TTS system assessment also. Our questionnaire (in Table 6) included three questions concerning intelligibility, general quality and naturalness.

A total of 1000 sentences were generated using sentences similar to those in [5] as well as narrative sentences. From these, 14 sentences were randomly chosen. One major problem in unit selection speech synthesizers is to make them produce robust quality. An all-inclusive or even a comprehensive evaluation through listening experiments is extremely difficult or impossible. The random selection was inevitable due to the varying quality of the sentences. Thus, we could have chosen a set of 14 sentences that would get the highest scores as well as choosing a set of 14 sentences that would obtain low scores at least for quality and naturalness. In addition, one real sentence from the database was added to the test set.

These 14 sentences were rated by 8 native Finnish listeners and the averaged results and the corresponding values of standard deviation are shown in Table 5. The ratings from different MOS tests can not be compared with each other, but an interested reader is referred to [5] where some commercial Finnish TTS systems were evaluated.

Table 5: Listening test ratings.

	average	worst sentence	best sentence
intelligibility	3.61 (1.00)	2.25 (0.89)	4.63 (0.52)
naturalness	3.00 (0.89)	2.38 (0.74)	4.25 (0.46)
quality	3.20 (0.70)	2.63 (0.52)	3.88 (0.64)

Table 6: Evaluation questionnaire.

INTELLIGIBILITY: did you understand everything without an effort, how would you describe the pronunciation?	
5	Excellent (no efforts, very clear pronunciation)
4	Good (small mistakes on pronunciation but did not bother)
3	Fair (a little annoying mistakes appeared)
2	Poor (annoying errors)
1	Bad (I did not understand the content because of too strong errors)
QUALITY: how would you describe the speech quality?	
5	Excellent (nothing bothered)
4	Good
3	Fair
2	Poor
1	Bad (I could not listen to speech of this quality a moment longer)
PROSODY AND NATURALNESS: did the sample sound natural?	
5	Very natural
4	Natural
3	Somewhat natural
2	Unnatural
1	Highly unnatural

6. Findings and future work

The database consisted only of approximately 1.5 hours of speech. CMU Arctic databases are of similar size but they were found too small in Blizzard Challenge [9]. It is a small database compared to the many commercial systems that use around tens of hours of speech. Due to the small database and its expressive style, naturalness and concatenation smoothness turned out to be somewhat contradictory requirements. The synthesis was found to sound rich in prosody but sometimes at the cost of concatenation smoothness.

The current TUT_VOICE system was implemented as a prototype and no extensive tuning of the system was done. The weights for the costs were tuned by hand but automatic phone-specific subcost training will be carried out in the future. Some very bad labeling mistakes were corrected manually and extra logic was included in the synthesizer to reduce label mismatches but finally the whole database should be manually corrected. In its current form, TUT_VOICE is not yet suitable for real-time speech synthesis but in the future, it will be modified to work real-time.

Creaky endings that are common in Finnish require some extra handling. The database design process took sentence-final syllables into account and did not accept them for coverage optimization. In the recordings, although special attention was paid on using creaky voice at the end of a sentence, they still appeared. In synthesis, creaky endings are currently handled through target costs, i.e. by penalizing the use of sentence final diphones elsewhere. On the other hand, the use of creaky voice quality at the end of a synthesized sentence can improve naturalness.

Sentences for the listening test were picked randomly and the results illustrated the general problem of unit selection: quality variability. The prosody of the synthesized speech was not rated as high as we expected. We found that it was mainly

because of strange phoneme durations. In Finnish, phoneme durations plays relatively important role. On the contrary, intonation is generally rather monotonic compared to many other languages. F_0 in the synthesized speech was found to be quite successful.

7. Conclusions

This paper described the design and implementation of a Finnish unit selection TTS system called TUT_VOICE. The quality of current commercial English unit selection speech synthesizers is high and the focus has moved into flexibility, for example generating new styles and emotions. The quality of TUT_VOICE is not yet at the same level, and one reason for that is a rather small database (1.5 hours). However, TUT_VOICE is a step towards natural, and style-variable flexible high-quality speech synthesis in Finnish.

8. Acknowledgments

This work was supported by the Academy of Finland, project No. 5213462 (Finnish Centre of Excellence program 2006 - 2011).

9. References

- [1] Wiik, K., "Fonetiikan perusteet", WSOY, Helsinki, 2nd edition 1998. 133 p.
- [2] Kominek J., and Black, A. W. "CMU ARCTIC databases for speech synthesis ver. 0.95", Language Technologies Institute, School of Computer Science, Carnegie Mellon University 2003, Available at <http://www.cs.cmu.edu/~awb/>. Referred 14.05.2007
- [3] Black, A. W., Taylor, P. A., and Caley, R., "The Festival Speech Synthesis System: System Documentation, Edition 1.4, for Festival Version 1.4.3", Human Communication Research Centre, University of Edinburgh, Scotland, UK. Available at http://festvox.org/docs/manual-1.4.3/festival_toc.html. Referred 14.05.2007
- [4] Project Gutenberg, Available at <http://www.gutenberg.org/>. Referred 14.05.2007
- [5] Ojala, T., "Auditory quality evaluation of present Finnish text-to-speech systems", Helsinki University of Technology, 2006, 65 p.
- [6] Vepa, J. and King, S., "Subjective Evaluation of Join Cost and Smoothing Methods for Unit Selection Speech Synthesis", IEEE Transactions on audio, speech, and language processing, 14(5):1763–1771, 2006
- [7] Lieko, A. "Suomen kielenfonetiikkaa ja fonologiaa ulkomaalaisille", Oy Finn Lectura Ab, Loimaa, 1992, 197 p.
- [8] Vainio, M., Werner, S., Volk, N., Välikangas, J., and Järvikivi, J., "Finnish Speech Technology: A Multidisciplinary Project", 2006, Unofficial web page available at <http://www.ling.helsinki.fi/suopuhe/>. Referred 14.05.2007
- [9] Clark, R., Richmond, K. and King, S., "Multisyn voices from ARCTIC data for the Blizzard challenge", Proceedings of INTERSPEECH-2005, Lisbon, Portugal
- [10] Hunt, A. J. and Black, A. W., "Unit Selection in a Concatenative Speech Synthesis System Using a Large Speech Database", Proceedings of ICASSP'96, Atlanta, GA, USA
- [11] Kondoz, A. M., "Digital Speech, Coding for Low Bit Rate Communication Systems", John Wiley & Sons, Ltd, West Sussex, England, 2nd edition, 2004
- [12] Iivonen, A., "Creaky voice as a prosodic feature in Finnish", Nordic Prosody, Proceedings of the IX Conference, Lund 2004, Frankfurt am Main/Berlin/Peter Lang, pp. 137–146, 2004.
- [13] Black, A. W. and Llitjos, A. F., "Unit selection without a phoneme set", Proc. of IEEE Workshop on Speech Synthesis, pp. 207-210, 2002
- [14] TUT_VOICE examples, Available at <http://www.cs.tut.fi/sgn/arg/ssw6/examples.html>. Referred 14.05.2007
- [15] Young, S., Everman, G., Gales, M., Hain, T., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., and Woodland, P., "The HTK book for HTK version 3.3", Cambridge University Engineering Department, 2005, 344p.