HELSINKI UNIVERSITY OF TECHNOLOGY
Department of Electrical and Communications Engineering
Laboratory of Acoustics and Audio Signal Processing

Tino Ojala

Auditory quality evaluation of present Finnish text-to-speech systems

Master's Thesis submitted in partial fulfillment of the requirements for the degree of Master of Science in Technology.

Espoo, June 6, 2006

Supervisor: Professor Unto K. Laine

HELSINKI UNIVERSITY OF TECHNOLOGY

ABSTRACT OF THE MASTER'S THESIS

Author:	Tino Ojala
Name of the thesis:	Auditory quality evaluation of present Finnish text-to-speech systems
Date:	June 6, 2006 **Number of pages:** 73
Department:	Electrical and Communications Engineering
Professorship:	S-89
Supervisor:	Prof. Unto K. Laine

Speech-producing computer systems have evolved so intelligent, that they fluently can read plain text input. Since these text-to-speech systems apparently have differences in the perceived sound quality, there is a need for research into the factors that affect the quality, and a need for the quantitative measurements of those factors.

Studies concerning synthetic speech have traditionally been conducted only for systems speaking languages of mainstream. In addition, there is only a limited amount of studies for the overall communicative capabilities of the systems, instead of concentrating into the details in speech production. In this work, the Finnish text-to-speech systems are evaluated for their sentence-level intelligibility in terms of "speech reception threshold" test, which was originally developed for testing the degree of hearing-impairment in humans. The test seeks for the speech presentation level that is barely intelligible in presence of noise.

"Speech reception threshold" test can effectively tell the difference in text-to-speech systems. A system, which produces sound from parameters, is found more intelligible over the systems that produce speech by concatenating pre-recorded speech samples. Reasons to this are the better spectral fit into the human hearing, smoother continuity of audio flow, less distortion and better possibilities for prosody modelling.

Keywords: Text-to-speech, Speech synthesis, Speech quality evaluation, Speech reception threshold, Hearing In Noise Test

TEKNILLINEN KORKEAKOULU DIPLOMITYÖN TIIVISTELMÄ

Tekijä:	Tino Ojala	
Työn nimi:	Nykyisten suomenkielisten tekstistä puheeksi -järjestelmien auditorisen laadun selvittäminen	
Päivämäärä:	6.6.2006	**Sivuja:** 73
Osasto:	Sähkö- ja tietoliikennetekniikka	
Professuuri:	S-89	
Työn valvoja:	Prof. Unto K. Laine	

Puhetta tuottavat tietokonejärjestelmät ovat kehittyneet niin eteviksi, että ne voivat lukea paljasta tekstisyötettä sujuvasti. Koska näillä *tekstistä puheeksi* -järjestelmillä kuitenkin mitä ilmeisimmin on eroja havaitussa äänenlaadussa, on tarvetta tutkia laatuun vaikuttavia tekijöitä ja saada kvantitatiivisia mittaustuloksia niistä.

Synteettisen puheen tutkimus on perinteisesti tehty valtavirran kielillä. Lisäksi sellaiset tutkimukset ovat harvinaisia, jotka selvittävät järjestelmien yleistä kyvykkyyttä kommunikaatioon sen sijaan, että keskittyisivät puheentuoton yksityiskohtiin. Tässä työssä suomenkielisten tekstistä puheeksi -järjestelmien lauseymmärrettävyyttä testataan *puheen ymmärrettävyyskynnys* -testillä, joka on alunperin tarkoitettu mittaamaan ihmisten kuulovamman astetta. Testissä etsitään sellaista puheen voimakkuustasoa, joka on juuri ja juuri ymmärrettävissä kohinan seasta.

"Puheen ymmärrettävyyskynnys"-testi pystyy tehokkaasti osoittamaan eron eri tekstistä puheeksi -järjestelmien välillä. Järjestelmä, joka tuottaa puhetta parametreista, paljastuu ymmärrettävämmäksi kuin järjestelmät, jotka tuottavat puhetta liittämällä ennalta äänitettyjä puhenäytteitä yhteen. Syinä tähän ovat parempi spektrisovitus kuuloon, juohevampi äänivirta, pienempi särö ja paremmat mahdollisuudet prosodian mallintamiseen.

Avainsanat: Tekstistä puheeksi, Puhesynteesi, Puheen laadun arviointi, Puheen havaitsemiskynnys

Acknowledgements

Hail to all who volunteered to the listening tests! Thank you, without you this work wouldn't have been possible. Hope you enjoyed the movie!

The following individuals I salute. Prof. Martti Vainio for sharing intellectual capital, SRT ideas, testing framework in University of Helsinki, and more. Prof. Unto Laine for the remote supervision from the land of Vespas' and for valuable feedback. Hanna Järveläinen for sacrificing time for guidance in subjective test problems. Janne Argillander for the initial ideas for the thesis and miscellaneous help along the way. Thank you all!

I also would like to thank all the companies who without prejudices offered their speaking machines into the demanding tests. An additional gratitude goes to IBM Finland for understanding the needs for this work and financially supporting me doing it.

Finally, I must thank everyone who showed interest in what I was doing during these months.

Otaniemi, June 6th, 2006

Tino Ojala

Contents

Abbreviations

ANOVA	ANalysis Of VAriances
BKB	Bamford-Kowal-Bench sentences
C	Consonant
CLID	Cluster Identification test
CTTS	Concatenative Text-To-Speech
DMCT	Diagnostic Medial Consonant Test
DRT	Diagnostic Rhyme Test
F0	Fundamental frequency of speech
F*n*	Formant *n*
HINT	Hearing In Noise Test
HMM	Hidden Markov Model
LP	Linear Prediction
MLSA	Mel Log Spectrum Approximation
MOS	Mean Opinion Score
MRT	Modified Rhyme Test
PB	Phonetically Balanced word lists
PH	high predictability sentences
PL	low predictability sentences
PSOLA	Pitch Synchronous Overlap Add
SNR	Signal to Noise Ratio
SPIN	Speech Perception In Noise
SRT	Speech Reception Threshold
sSRT	sentence Speech Reception Threshold
ST	Short-Term
TTS	Text-To-Speech
V	Vowel

List of Figures

List of Tables

Chapter 1

Introduction

The industrial revolution has shown that it is easy to let the machines do the work that always repeats the same, so that the workers can be released to do jobs that are more demanding. At first, the tasks were about easy mechanical automation, but since, always more and more difficult problems are left for machines to solve. One of the most common tasks human do is communication by voice, which also can be seen as consecutive repetitive voices from the mouth, although the underlying messages may vary. Therefore, the speech is also an obvious task for a machine to accomplish.

One ultimate goal for speech producing machines is presented in the sci-fi story "2001 Space Odyssey", where the speaking computer, Hal, controls the spaceship and meanwhile chats fluently and intelligently with the passengers. In the movie version from the 60's, Hal had to be played by a voice actor. Although the date of the plot has passed, there is still no machine equal to Hal in speech naturalness, because in most cases, people can tell the difference between natural and artificial speech.

This thesis first discusses the factors that affect the speech quality and the current situation in artificial speech research. Secondly, and more importantly, it sets the Finnish text-reading machines under a quality test.

1.1 Structure of thesis

Thesis is structured as follows. The rest of the chapter briefly introduces the linguistic concepts and terminology needed to follow the rest of the work. It also describes the basics of human speech production that is referred later on in the speech synthesis methods. Chap. 2 describes the technology that is used in intentional production of artificial sounds that are recognized as speech by human listeners. Chap. 3 discusses the quality aspects of speech, and the measurement methods especially used to evaluate the quality of synthetic

speech. The evolution of intelligibility tests for hearing impaired are also presented with a description about how the ideas are used in the quality evaluation of synthetic speech in this work. The evaluation process is described in Chap. 4, and the results are examined and discussed in Chap. 5. Chap. 6 summarizes everything that has been done, discusses how the work benefits the speech research, what were the deficiencies, and how should those be corrected in the future. And of course, the auditory quality of present Finnish text-to-speech systems will be uncovered.

1.2 Linguistic concepts

1.2.1 Speech production

Human speech is a complex combination of sounds generated by vocal organs (see Fig. 1.1). The most important organs in all speech are the lungs combined with the muscles, such as diaphragm, surrounding them. Together they provide the upper organs the air pressure, which is the main form of energy in the speech. The source of voiced sounds is the vocal cords, which vibrate when air from the lungs is flowing through them and they are tensed by the surrounding muscles. The vibration rapidly cuts the airflow; when the vocal cords are closed, almost no air is let through and with respect to the cords opening, the volume velocity of air increases. The fundamental frequency of the pressure variation can be examined from the frequency of cords closure, being around 100-200 Hz depending on person. The area between the vocal cords is called glottis, and hence the sound source is often called the glottal source.

Pressure variation from the glottis does not transfer out as it is, but it is highly modified by the cavities in other organs, that produce the *articulation*. Vocal cords are connected to pharynx cavity, which opens to oral and nasal cavities, from which the sound flows out through mouth and nose. The route from vocal cords to mouth is also called vocal tract and it is the most important modifier of articulation. By changing the shape of vocal tract with vocal organs, different combinations of resonances will occur. The nasal cavity has fixed dimensions but it affects the total resonances when *velum* lets air to flow through it.

Unvoiced sounds are produced without the help of vocal cords. Such sounds arise from constriction in vocal organs so that the airflow gets turbulent and noisy in sound. If vocal tract totally closes and rapidly opens, plosive sounds are produced. Only a very few sounds can be made using bare vocal organs without even the airflow from the lungs, mostly pops with lips and tongue.

Vocal tract can be considered as an adjustable acoustic filter. The glottal excitation has complex frequency content with lots of harmonic frequencies in addition to the fundamental. The resonances in vocal tract emphasize corresponding frequencies of glottal excitation.

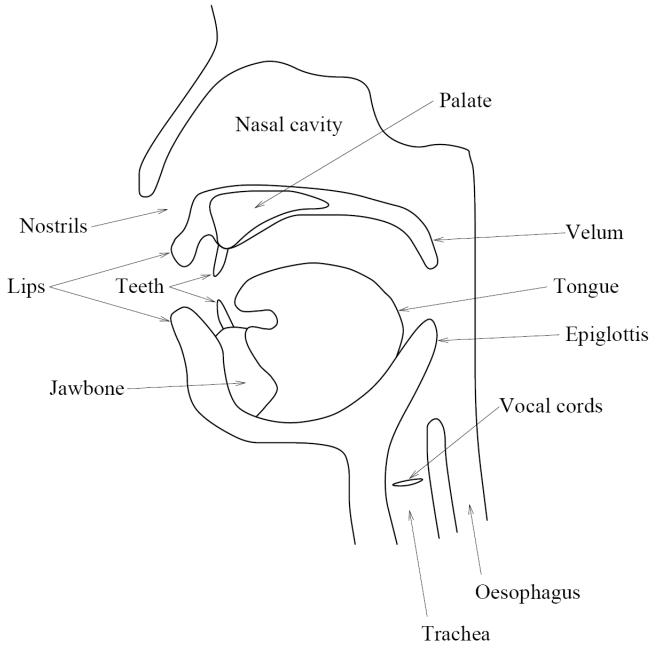

Figure 1.1: Most important vocal organs.

These peaks are called *formants*, which are labelled as $F1$ for the first harmonic (formant 1), $F2$ for the second harmonic (formant 2) and so on. Correspondingly, the fundamental frequency of the glottal excitation is labelled as $F0$.

1.2.2 Phonetics

Linguistics believe that a *phoneme* is an atomic unit of human speech, an abstract class of several similar perceived speech sounds. Phonemes are often conceived equal to the written letters, but this is not the case in general, since written and spoken language differs, and the same letter may be pronounced differently regarding the context. The phoneme perception involves cognitive processes like knowledge of the language; if some phonemes do not exist in a language, they are hard to percept and pronounce correctly. For example, Swedish word "hus" has a phoneme that is an intermediate form of Finnish /u/ and /y/ and the word is likely heard as /hu:s/ or /hy:s/ by Finns. As one letter may cause several phonemes, also a phoneme can be extracted from objectively different kinds of sounds. For example, in phoneme combinations /uku/ and /iki/, the vocal organs are in different shape during the speech. It results in different sounding phoneme /k/, which is nevertheless percept as the same. In English, there are about 40 phonemes (see Donovan 1996) and in Finnish, 24 (see Karjalainen 1999).

Corresponding acoustic realization of a phoneme is called a *phone*. A phone is thus the sound humans perceive and which can be categorized into the class of corresponding phoneme. Different-sounding phones belonging to the same class of a phoneme are called *allophones* (that is, /uku/ and /iki/ have the same phoneme /k/, but two different allophones of /k/).

As phoneme is an atomic unit of spoken language, an atomic unit of written language is called *grapheme*. In principle, graphemes are the letters of a language extended with other symbols like punctuation marks and numerals. An *orthography* is the set of symbols used when writing a language. For example in Finnish orthography, a letter corresponds to a grapheme, and in phonological orthography, a phoneme corresponds a grapheme.

Basic division of the phonetic alphabet is between vowels and consonants. Vowels are relatively stable, open voiced sounds. They are produced by relatively stationary periodical glottal excitation combined to filtration of fixed vocal tract. Characteristic to those is that they can be extended to last as long as there is excitation (/a/ . . . /aaaa/), or in practise, as long as it takes to empty the lungs. Stability means that there can be found a constant pitch: for example the melodies in songs are based on varying the pitches of vowels. Traditionally, the Finnish vowels are classified like in Figure 1.2. The leftmost column characteristic is the mouth opening and the second is the tongue position (high-low). Uppermost row characteristic is the tongue position (front-back), and below that, there is the vocal tract constriction shape.

Vowels		front		back	
		wide	round	wide	round
Close	high	i	y		u
Close-mid Open-mid	mid	e	ö		o
Open	low	ä		a	

Figure 1.2: Finnish vowels sorted by mouth opening (close-open), tongue position (high-low, front-back), and vocal tract constriction shape (wide-round)

Phonemes that are not vowels, are consonants, which can be voiced or unvoiced and may have changing sound characteristics during their appearance. They are usually in contact with vowels or used to connect them. A traditional classification of the Finnish consonants is presented in Figure 1.3. *Plosives* (or stop consonants) are those that do not need lung-produced airflow, but can be articulated with bare upper vocal organs. *Fricatives* are

constant, turbulent, and often non-voiced sounds not having a clear pitch. Voiced fricatives also exist, such as voiced /z/ in some languages. *Nasals* arise from voiced sound flowing through nasal cavity combined with the resonances in the vocal tract. The opening of the vocal tract after nasal cavity usage also plays an important role in nasal production. *Tremulant* /r/ is considerably different from other phonemes. A rattling sound is produced by tongue tip vibrating in resonance against the upper wall of the vocal tract. /r/ is normally connected to vowels, making it a voiced sound. *Lateral* /l/ is a voiced sound trough vocal tract from beside the tongue. *Semivowels* are used to modify adjacent vowel continuation.

Consonants	labial		dental alveoral			palatal	velar	laryng.
	bi-lab.	labio-dent.	pro	medio	post			
plosive (tenuis)	p		t				k	
plosive (media)	b			d			g	
fricative (sibilants)			s					
fricative (spirants)		f						h
nasal	m		n					
tremulant			r					
lateral			l					
semivowel		v				j		

Figure 1.3: Finnish consonants sorted by their characteristics. See text for more information.

1.2.3 Prosody

Prosody refers to intonation, rhythm, and vocal stress in speech, and especially their variations. Intonation, in general, refers to the pitch of the sound. In music, intonation means the accuracy of pitch being intended, and in speech, it means the pitch variation during pronunciation of utterance. Languages are spoken in learned intonation and they can be shaped intentionally to emphasize something in the message. For example, in French the pitch rises when asking "Ça va", but falls when the same is said as answer.

Rhythm of the speech is the timing of units such as emphasized syllables and breaks between words. Variation of speed in speech can also be regarded as a rhythmic feature.

Vocal stress is a general term of making some parts of speech more emphasized than other parts. The most obvious, and hence sort of a definitive way to do this, is to increase loudness. Intonation, rhythm, and tone changes may give similar emphasizing effect, so the term "stress" can mean whatever emphasis of speech, or when exactly defined, only the

emphasis by intensity.

Speech units of all lengths have prosodic features. Words commonly have main stress on one syllable and if the word is long, it may contain two or more stressed parts. The pitches of different vowels in a word vary. At sentence level, speech has lively intonation, some stressed words, and rhythmic variations, like simply pauses between words. There is usually a longer pause between two sentences while the speaker draws breath. At paragraph level, when reading text, there can be found further prosodic features. Whole sentences can be stressed if they are found more important. Because a new paragraph is started to describe something different from previous, also a longer pause is required to distinguish from a sentence pause.

1.2.4 Other linguistic concepts

Syntax of the language is the rules how the words can be combined. It concerns how different words, such as verbs, nouns and adjectives are aligned to form the sentences in a language. *Semantics* refers to the meaning of the sentence, whether the sentence is reasonable. It is an abstract concept for human interpretation of the language. The difference between syntactic and semantic concepts can be seen from a sentence that is syntactically correct, but semantically incorrect, such as "Red weather found the cheese."

Chapter 2

Speech synthesis

Text-to-speech (TTS) is a common term for a system that converts any written text into audible speech. The task has been evolved highly computational, so that the solutions are generally made to accomplish with a general-purpose computer instead of strictly dedicated components. Due to this, the TTS term is widely accepted to cover only text that is formatted into a form that a computer can read, usually ASCII text files. An area, which is interesting but usually out of scope of TTS discussions, is the conversion of any form of text (for example hand-written) via text files to artificial speech. At the other end, the production of speech needs equipment to convert binary audio representation into audible sound. This can also be out of interest, because it is rather easy to switch between, for example, loudspeakers, headphones, and telephone headset, that all give different sound. In many cases, it is enough for TTS system to produce sound files for other sound producing systems. At some level however, the sound production has to be present. It can be disputed whether there is speech although nobody is hearing it, but there surely has to be sound for speech to exist. Altogether, the term TTS can be defined as a use case: "Feed in text and it will give out speech".

TTS definition limits out the simplest systems that use pre-recorded speech samples and collects continuous speech from those. These systems are well known from telephone voice-response systems or public announcements at buss stations; the message "*Bus to Lahti leaves at 12:15*" is automatically generated by selecting the correct words from among all recorded words. In these type systems, a few ten to a few hundred words or phrases are recorded to cover all situations needed, for example, to announce the bus schedules. All the new applications that require different vocabulary need new recordings. However, TTS is expected to speak any text, so word-collection systems cannot be regarded as TTS. It is impossible to store arbitrary words in all forms of inflection as discrete recordings. Besides, it is desirable that the machine could speak in natural-sounding manner, instead of just mono-

tonically repeating a word after another, which would be the result of simply concatenating the words in arbitrary order.

Task of converting text to speech can roughly be split into two phases, like in Figure 2.1. In-fed text is first analyzed and it is converted into linguistic representation. In simplest form, this has similarities to phonetics in a dictionary, which tells how the word is pronounced. However, when continuous text is converted, the prosody has to be taken into account. This phase also differs in challenge between languages: in Finnish, the most of the language is spoken similarly as written, but in other languages, there can be plenty of exceptions in pronunciation rules. In the second phase, the linguistic representation is synthesized to audible speech. It contains all the sound processing, like speech "melody" - the rising pitch of question phrases etc.

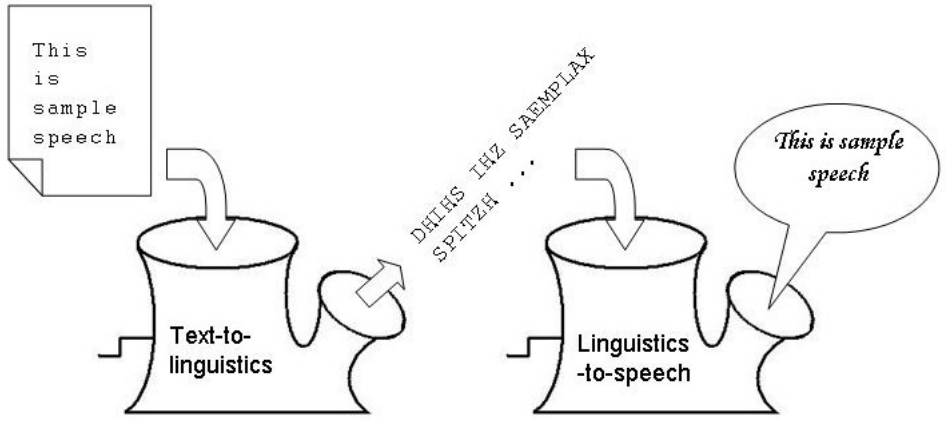

Figure 2.1: Two main phases in TTS. Text is first converted into linguistic representation, which is then converted to audible speech.

It is worth to notice that the term "speech synthesis" is extensive and it is used to describe plain phonetics-to-acoustics conversion as well as the whole framework of generating speech by a machine. For coherency of terminology, it is better to separate TTS systems from bare synthesis systems, which can also be driven with other kinds of inputs instead of text, for example with predefined rules. In literature, also terms "high-level" and "low-level" synthesis are commonly used to distinguish between text-to-linguistics and linguistics-to-speech tasks, respectively.

The following sections will discuss about the two phases of TTS, concentrating more on low-level synthesis, since it apparently has more deficiencies and solution concepts.

2.1 Text-to-linguistics

(Klatt 1987) introduces a good description about text-to-phonemes conversion. The task is divided into the following sub-tasks where a sentence of ASCII string is fed in and a linguistic presentation of the string is given out.

- Reformat everything encountered
- Parse the sentence to establish the surface syntactic structure
- Find the semantically determined locations of contrastive and emphasis stress
- Derive a phonemic representation for each word
- Assign a (lexical) stress pattern to each word

These five points are briefly examined below. In addition, the first points are clarified by parsing an ASCII string with a fictitious, very basic text-to-linguistics converter. Let the example phrase be "*I - like other boys - have been working for 12 hours!*"

ASCII string can contain characters that are not pronounceable, such as punctuation, digits, special characters, and abbreviations. Basic punctuation is reasonable to left in place for later stages to find out the boundaries of the sentences, but others are to be reformatted as fully written text. For example *eur2.20* have to be changed to something like *two euros and twenty cents*. Easiest way to do this is to use table lookup for replacing substrings in issue, but there has to be clever rules to do that, because abbreviations most likely have different meaning in different contexts. For example, *RIP* should be converted to *Routing Information Protocol* when talking about data networks, *Rest In Peace* as a benediction for the dead, but it might also be a standalone word. It is reasonable that the lookup tables may be switched regarding to the usage of TTS system, so that the abbreviations will match the wanted vocabulary. The example phrase becomes *I - like other boys - have been working for twelve hours!* (not *I minus other boys...*).

Syntactic and semantic parsing helps to analyze sentence's pronunciation in entity. Generally, language grammars conduct some syntactic rules how the words are to be ordered in sentences. This affects the prosody of the spoken sentence. Phrase boundaries are easily found from punctuation and conjunctions so that e.g. lower stress of subordinate clause can be produced. Other examples of items affecting phrase prosodic boundary are the speech parts such as verbs, adjectives, and prepositions. Verbs need to be recognized, because they most likely get the main stress in the sentence. Additional difficulty in this is the words that can be interpreted as a noun or a verb, so the intra-word stress can be different - word *permit* has stress on first syllable as verb but on last as noun.

Prosodic rules are placed among the ASCII string in a way the low-level synthesizer can later on understand them. In the example phrase, a siding sentence is detected from the dashes and is given less emphasis. Otherwise the phrase would start "*I like other boys...*" and henceforth having confusing verbs. The last word of a sentence decreases in pitch. The verb *working* receives the main emphasis. The exclamation point at the end makes a few last words stressed. Let the formatting be preceded by "@" mark: *I @pause() @faster@low_stress@low_pitch(like other boys) @pause() have been @slower@high_stress@high_pitch(working) for @high_stress(twelve) @decreasing_pitch@high_stress(hours)*

Semantic analysis takes the meaning of the sentence into account when forming the prosody. "*An old man sat in a rocking chair*" has several different ways to be pronounced. If the age of the sitter is important, *old* has to be stressed. Humans know from the previous knowledge that he is sitting on something called *rocking chair*, but poor analysis could emphasize *rocking* as an adjective. At present times, this kind of meaning-dependent and pragmatic formatting is extremely difficult computationally, and most likely introduces problems in sentences with non-standard pronunciation.

Also being a semantic concept, the emotional effect can be added to voice characteristics. The example phrase seems to be said in anger or in frustration. It makes the sentence to be spoken in increased intensity and dynamic changes (see Lemmetty 1999, Chap. 5). If parser could recognise anger from the text, it could insert corresponding prosody tags to the string. Again, the semantics are not easily implemented.

Text-to-phonological orthography -conversion is quite straightforward in Finnish where the language is mostly spoken as it is written. This is not the case in general, like in English, where the written form of the text differs from the spoken. It is also common that the same base form of a word is pronounced in different manner depending on affixes, as in words *sign* - *signal*. At least three base methods of converting text to phonetics in word level can be considered: rule based system, pattern learning, and morphemic decomposition.

Rule based letter-to-phonemes systems try explicitly put to effect languages' phonological rules, which they, to some extent, have. A challenging thing is to find correct amount of letters in a word to be parsed at time, so that the correct pronunciation can be achieved. There can be over 500 rules in these systems, like say the letter *a* as */e/* if followed by *ve*. The rule works for *behave*, but nor for *have*, so more sophisticated rule becomes apparent.

Pattern learning has similarities as human learning. If a new word is encountered, one can think: "This new word is almost the same as an older that I know. Let me pronounce it likely!" With large amount of speech and corresponding pronunciation data, computer can also be trained for statistical recognition of pronunciations or to use analogy from similar, known words, and that way learn new words.

In morpheme decomposition, the words are broken into smallest possible meaningful parts, morphemes, whose pronunciation are then retrieved from storage, or predicted. Breaking is done by e.g. removing affixes and splitting compounds. Word "hothouse" is split "hot" and "house" to get rid of letter pair "th", which is otherwise pronounced like in "without". Advantages in this technique are efficiency of presenting over 100000 English words with 12000 morphemes, and the ability of helping syntactic analysis to specify speech parts. This method seems to be the most powerful, having only a few percent error rates in random text.

All three text-to-phonological ortography methods above will benefit from using a pronunciation dictionary, where the most common words' pronunciations are stored. With a dictionary of 2000 words, already over 70% of random English text can be converted, so there is noticeably less possibilities for algorithm-dependent errors.

Lexical stress tells how the syllables in a word are to be stressed. The stress can be predicted with similar approaches than text-to-phonemes conversions: using rules or morphology. Syntactic analysis may also affect this phase. Assigning lexical stress is perhaps the weakest link of all text-to-linguistics conversions. Incorrect stress patterns are disruptive to listeners and may even trigger miss-selection of vowel qualities in later stages.

2.2 Current trends in linguistics-to-speech

Throughout the history of speech synthesis, there have been several different methods to produce speech-alike audio signals, starting from mechanical copies of human speech production anatomy and ending to complete mathematical models of everything involved. Only current trends are presented in this thesis; a good overall review of the issue can be found in (Lemmetty 1999).

Linguistic-to-speech, or low-level, synthesis can be categorized into three methods based on how the resulting speech waveform is produced. These categories are concatenative, parametric, and articulatory synthesis, which are all briefly described below.

2.2.1 Concatenation synthesis

Term concatenation refers to connecting items after another. In TTS terminology, the concatenated items are pre-recorded units of speech, together forming understandable speech from arbitrary text. The easiest way to reproduce natural-sounding speech is to record a predefined speech passage and replay the same speech. The spoken passage can be reformat to have new meaning by cutting parts of it and pasting them in different order, or in other words, concatenating the parts. All the concatenation points, however, are likely to incorporate discontinuities in speech. The longer the units are, the less concatenation points

and possible discontinuities occur. The length of appropriate concatenation unit has been a matter of discussion throughout the history of concatenative TTS (CTTS).

As concluded at the beginning of this chapter, concatenating recorded words is not a solution, because unlimited amount of characters strings, namely words, can be formed with limited number of characters. It yields that there is not enough storage capacity for all units in the whole world! This is partly true, since new words, proper names etc. will appear continuously.

Syllables, and their halves, demi-syllables are found when words are broken into parts. These provide already more reasonable solution for concatenation synthesis. Almost everything in a language can be spoken with a limited number of syllable-sized units (Donovan 1996). The storage/memory needs are still high, but growing capacities may provide a solution. The bigger issue in these are the co-articulation problems, which arise from difficult smoothing of boundaries between units. Units of these sizes are not widely used, as even smaller are preferred.

A phoneme was found to be an atomic unit of speech in section 1.2.2. A good example of leashing a psychological concept into pragmatic use is that of successful use of phone and corresponding length units in concatenation synthesis. One of the most popular units in CTTS systems are diphones. A diphone is a unit that consists of the last half of one phone followed by the first half of another. The benefit of diphones is the smooth transition between phones, which would otherwise be difficult to achieve. Because of allophones, the boundaries of a phone may appear different regarding the context, but the middle state of the phone usually appears quite steady. This steady state of phone is a reasonable location to cut a part of speech, and to paste in another part. In principle, the amount of diphones needed is square of all possible phones including allophones (Lemmetty 1999), but some of those that are not present in a language, can be neglected. This yields a few thousand diphones, which is not a storage issue of any kind. The problems may potentially arise when two abutting diphones do not reach the same vowel target (Klatt 1987).

Phones themselves are disadvantageous in concatenation synthesis, because of the context dependency described above. However, the evolution of automatic determination of phone boundary location, and the use of extensive amount of allophones have made this unit size also suitable.

Units shorter than phones, sub-phones, are advantageous because speech in general becomes more acoustically self-similar in such time scales (Donovan 1996). A sub-phone can be presented with a vector containing only the features from which the waveform can be reformed, instead of presenting the waveform itself. Both sub-phone and phone presentations of speech contain large amount of context dependent variation, so there has to be large amount of those segments available to produce unrestricted speech.

An essential part of concatenation synthesis is a carefully prepared acoustic inventory from which the segments are retrieved for concatenation. All the possible units (e.g. diphones) need to be recorded into storage, and they have to be segmented so that their boundaries are cut from suitable locations. An inventory can be achieved by recording predefined words (perhaps of nonsense) that incorporate all the necessary units, or by recording natural passages of speech, from which the units are extracted. In the latter, it is important to take care of having all the units included in the passages.

Extracting parts from continuous, recorded speech, and concatenating them to form new speech, is called *unit selection* synthesis. For unit selection, the speech recording is usually done by reading some continuous text into storage. The recorded speech is also transcribed into written phonetic representation, which is used to help the segmentation of speech units, and labelling them. To get the segmentation accurate, the phonetic transcription has to be tightly coupled to the recorded speech in issue, instead of using some general linguistics of the language. After segmentation, each segment is provided a corresponding phonetic label regarding to the transcription, and in addition, a vector of several features, like pitch, amount of stress etc. can be attached to the segments. Construction of acoustic inventory benefits of automated procedures, but has traditionally been made by trained humans, which is slow and erroneous (Donovan 1996).

In unit selection TTS, the high-level synthesizer (see Sect. 2.1) provides a linguistic representation of the text to be synthesized. The acoustic inventory is searched for such units, that the concatenation of them results best match to the target linguistics. The optimization of unit selection can be based on cost functions (Hunt & Black 1996). The idea is to select units with minimum the total cost, which is the sum of concatenation cost and target cost. Former is an estimate of smoothness in joint between consecutive units, and the latter is an estimate of difference between a unit and the target it is supposed to represent. Two units that were consecutive in the original speech, is a special case of concatenation cost, its being zero. Otherwise, the features of previously selected unit and the features of following candidate are compared and concatenation cost is calculated. Target cost is calculated comparing the features of desired unit in place and features that the candidate has.

Plain unit selection scheme can produce highly natural speech, but the disadvantages are distortion due to spectral discontinuities in concatenation joints, and poor prosody. To get the concatenation smooth, the recordings have to be made in relatively monotonic voice, so it introduces as small discontinuities as possible in the joints of speech segments. This yields the resulting speech lacking temporal, volume, and pitch variation, which means bad prosody modelling of rhythm, stress, and intonation, respectively. To get these features available, it requires that every unit appears in every prosodic context in the acoustic inventory. With a large speech database, a variety of occurrences of the same unit are available,

each appearing in different phonetic and prosodic content. The selection is done by minimizing the target cost also regarding the prosody, but still the whole utterance most likely will suffer from lack of it. As a solution, some signal processing tricks can be used for artificial shaping of the resulting speech. Adding the stress by adjusting the volume is a trivial signal processing task, but independent pitch and duration regulation needs approaches that are more sophisticated. One of common methods is *pitch synchronous overlap add (PSOLA)*, where the speech is broken into short-term (ST) signals. This is done by providing pitch markers to the voiced regions of speech, and centering a Hanning-type window on the mark. For unvoiced regions, the Hanning windows are located with fixed time intervals. Now, the duration and pitch of speech can be altered by multiplying or removing ST signals, and changing the spacing of them.

Phrase splicing (Donovan et al. 1999) represents a progressive unit-selection synthesis method. It also has a speech database, from which the concatenative units are selected. Instead of selecting constant, predefined units, phrase-splicing TTS tries to find as long speech segments as possible matching the text to be synthesized, from the database. A fortunate chance is to find a whole phrase of input text as recorded. The speech segments that are not found as entire phrases are produced with traditional unit-selection scheme discussed above. Phrase splicing naturally suffers less from the concatenation discontinuities than other unit-selection methods. The possibilities of finding long phrases in the database can be increased by anticipating the usage of the system under development, and recording probable phrases to the database. This kind of tailoring makes the phrase splicing system have similarities to those simple ones, which use pre-recorded sentences to produce announcements etc.

Fundamental, in a way unfavourable properties of concatenation synthesis are the speaker dependency and space requirements. With one speech database, only the voice of the person used in recording can be re-synthesized. If other voices are needed, they have to be prepared individually including all the stages of building an acoustic inventory. Each voice's inventory size is still considerable with current storage capacities and if context dependency is needed, the inventory will grow further. The sizes of inventories of unit-selection TTS systems can easily be several hundred megabytes. *Linear Prediction (LP)* is a common method in speech signal processing, and can be used in compression of speech database size. In addition to storing the LP-coded speech, some methods are introduced for a whole speech synthesis scheme using LP coefficients to parameterize the vocal tract and simplified residual to approximate the glottal excitation. However, at very simple scheme, this results in quality far from perfect (Klatt 1987).

2.2.2 Parametric synthesis

While concatenation-based synthesis uses pre-recorded speech waveforms, the parametric synthesizers create the waveforms themselves from specific parameters. They often take advantage of source-filter theory of speech production as in (Fant 1970), as drawn in Fig. 2.2 (see also Lemmetty 1999, chap 1). There are two sources of voice, from which the voiced source corresponds to the glottal source, and the unvoiced corresponds to noise-alike sounds in human speech production. The filter part approximates the vocal tract at some level, most commonly filtering formants. As time goes, the source of excitation, the gain, and the filter coefficients are adjusted following explicit rules, so that the system produces time-varying signal that sounds like speech.

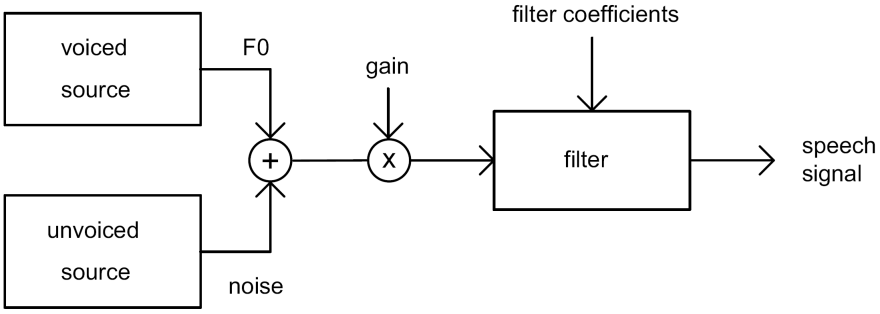

Figure 2.2: Source-filter model of speech production

In formant synthesis, the filter part of source-filter model is a combination of individual band-pass filters in frequency band of formants. The filters can be connected in cascade or in parallel. The cascaded type combination has been found better for non-nasal voiced sounds, while parallel type performs better in nasals, fricatives, and stop consonants. Therefore, complex combinations of both types has also been proposed having several filters in parallel and cascaded. (Klatt 1987) describes the evolution of voiced source from a simple filtered impulse train to more sophisticated and natural model. The unvoiced source is filtered noise, perhaps of white type.

In the past decades, a new approach of *Hidden Markov Models (HMM)* has been successfully brought into the area of speech synthesis, partly because of good performance in another speech processing task, speech recognition. An HMM is a statistical finite state automate that produces observations as output (see Fig 2.3 for a three-state left-to-right HMM). Every state has transition probabilities, by which the transition from current state to some other state occurs. Always, when a transition is taken, an observation is emitted regarding to the state's output distribution. The transition can happen from a state to whatever other state, but the current state cannot be seen straight from the observation, because

the same output can be the result of any state, although it can be more probable from certain states. The use of HMMs is that the parameters can be trained so that certain states are associated with certain features in their output distributions. More about theory of HMM-based speech synthesis can be found for example in (Masuko 2002).

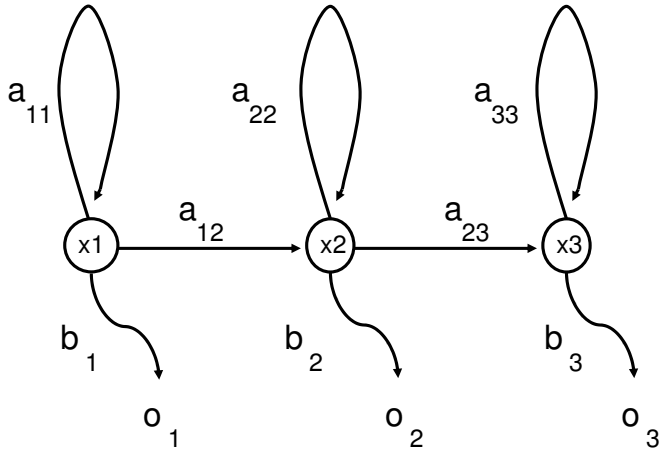

Figure 2.3: A three state left-to-right HMM with states x_x, transition probabilities a_{xx}, output probabilities b_x, and observations o_x

(Tokuda et al. 1995) used HMMs to model sequences of speech spectra. The spectral estimates were extracted from natural passages of speech with mel-cepstral analysis (Imai 1983) that produced 13 mel-cepstral coefficients for a speech frame of length 25.6 ms. From these coefficients, the speech can be re-synthesized directly using a "Mel Log Spectrum Approximation (MLSA)" as a filter in source-filter model. The coefficients were used as spectral parameters to train HMMs, which, when appropriately trained, should be able to reproduce the parameters. Only a very simple, unreal language with three phonemes /a/, /i/, and /o/ was assumed, so the parameterized versions of those were needed as the training data. Each of the three phonemes was assigned an own 3-state HMM. In synthesis phase, the HMMs emitted vectors of mel-cepstral coefficients as observations. For smooth spectral transitions between different observations, the dynamic features are incorporated into the mel-cepstral coefficients vectors (c) as first and second order differentials (Δc and $\Delta^2 c$) between successive vectors.

In principle, the HMM synthesis system concatenates the phone-length HMMs. The method is not included in the concatenation synthesis section, because it synthesizes speech with the source-filter model from the statistically trained parameters, instead of concatenating pre-recorded waveforms.

Structure of a feature vector that is used to train a HMM, and which is to be obtained

from it in the synthesis, is presented in Fig. 2.4 (Yoshimura et al. 1999). It also contains the excitation part in addition to the spectral part for the both parts to be treated simultaneously. Each state transition in an HMM emits a vector of this kind as an observation, and because the state transition can happen back to the same state, similar consecutive vectors are probable.

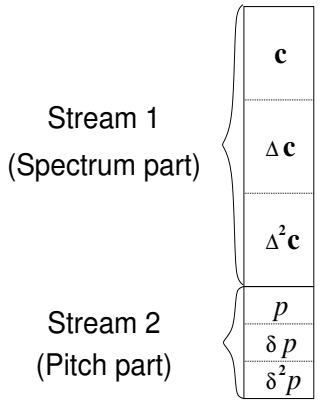

Figure 2.4: Feature vector

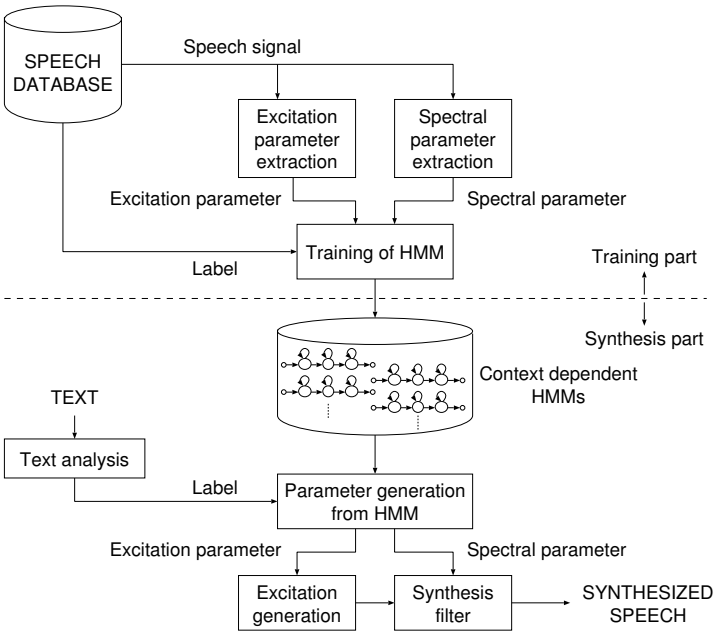

Figure 2.5: HMM-based speech synthesis system

The complete HMM-based speech synthesis system (HTS) can be seen in Fig. 2.5 as presented by (Tokuda et al. 2002). In the training part, the recorded speech signal is broken into spectral part (mel-cepstral coefficients) and excitation part (F0). Phone-length context-dependent HMMs are trained so that it results in having models of all the phones in as many prosodic contexts as possible. The label for each HMM is retrieved from the speech database. The HMM models are clustered in decision-trees regarding to several contexts, e.g. models with similar spectrum are found in the same cluster, as models with similar F0. This overcomes the problem, that it is impossible to have all the combinations of contextual factors in the training speech. In the synthesis part, the text to be synthesized is first converted to a context-based linguistic representation by a high-level synthesizer. Context dependent HMMs corresponding to the linguistic labels are retrieved from the inventory, and concatenated. The observations from HMMs are feature vectors of type presented in Fig. 2.4. From the observation spectrum and F0, the speech waveform is synthesized directly using MLSA filter.

HTS produced speech is smooth and stable compared to unit selection, which suffers from discontinuities in concatenation points and possible miss-selection of units. In addition, the data size of an HTS engine can be below 1 MB, and it can incorporate several voices. A disadvantage is the naturality: HTS speech sounds vocoded and buzzy (Tokuda et al. 2002).

2.2.3 Articulatory synthesis

The most computationally form of speech synthesis is the articulatory synthesis, where the air behaviour is directly computed from the mathematical models of vocal organs. The construction of articulatory synthesis involves measurements of human vocal organs in action, and development of corresponding geometric models of vocal tract and possibly models of glottal source.

The difficulty in articulatory synthesis is the data acquisition. Measurement equipment should not disturb the articulators' movements, so the wired intra-oral sensors need careful design. Remote imaging has own advantages and problems, such as the x-rays producing only two-dimensional pictures and the "magnetic resonance imaging" being too slow to capture fast movements. In addition, running the model of moving vocal tract requires high computational power. For these reasons, the articulatory synthesis is still used mainly in basic research, and there are no complete TTS systems available. A review of current situation of articulatory synthesis can be found in (Palo 2006).

Chapter 3

Speech quality evaluation

Quality is a relative concept that is most likely misunderstood if there is not a definition that is accurate enough. Quality is used in comparisons of similar items when there are no quantitative features available, or they are otherwise difficult to use or unwanted. If an item is in some sense better than other is, it is said to have better quality at the point of issue. Quality cannot be assigned if there is nothing to compare, or there exists only one item in the area of interest. When comparing items in terms of quality, there always is a way to judge the best quality to whichever item, by modifying the definition of good quality towards the properties of selected item. One item might be easiest to use in the group and that way have the best quality, while other might have the best quality by being most colourful. For this reason, when studying the quality, the definition of what is good has to be set first.

Auditory quality is limiting the definition of quality to the features that can be heard. The items under comparison have to make sound for them to be compared qualitatively. This restriction still leaves unlimited amount of features in place, one item might have the loudest sound, and the other might have most pleasant, both of which can be regarded as better quality. The term "auditory quality" concerns the human perception of sound, and suggests that the qualitative comparison is made by listeners who do judgements according to their own preference of the best quality. Measurement equipment can be used in some cases, if there is knowledge about correspondence between measured quantity and perceived auditory quality.

Qualitative comparison made by humans giving their opinions about comparable items, is called *subjective*. Subjective quality can differ between people; some may find the same thing as better quality than others do. There is no correct way to define good subjective quality, only some conclusions can be drawn about what is generally accepted. The converse of subjective is *objective*, which is not dependent on individuals opinions, but can be measured similarly by anyone.

Quality evaluations are often done by comparing pre-selected features in several items. The items can be ranked in order by comparing them among their group, or there can be a reference that is somehow of known quality and the items are compared to that reference. Sound quality assessment of a transmission channel can be made by playing the original sound as a reference, and then playing the sound through channel under test. Subjective opinions - as well as objective measurements - about the quality of transmission channel can be studied on basis of the differences between the reference and the tested sound. When comparing speech synthesis systems, however, the "original" sound is created by the system under test, so there is no possibility of straight objective comparisons of the quality. The ultimate reference is always an individual image of speech and how should it sound like. The synthetic speech quality evaluation is a subjective task where the subject judges what he hears against what he appraises as perfect speech. Comparisons between features of different systems can be made, and the system with features that are closest to the individual's intuition about perfect speech, will have the best quality.

Features, that are reasonable to assign quality in synthetic speech, are intelligibility, reading comprehension, naturalness and suitability for particular application (Klatt 1987). Good quality in intelligibility means that all the spoken utterances can be understood by the listener, including correct perception of units of all lengths, from phonemes to sentences, or longer. Comprehension, being similar concept as intelligibility, is used with overall understanding of the spoken language. However, the border between concepts of intelligibility and comprehension is inconstant. The difference of those could be thought as follows: The speech may have good quality in intelligibility of short units, but still it may be difficult to comprehend the message in whole, when listening to long stories. A reason for this could be bad prosody, so that individual sentences are easy to understand, but interconnected sentences do not seem to belong together and that makes the speech uneasy to follow.

3.1 Traditional measures

3.1.1 Segmental intelligibility

Synthetic speech is traditionally evaluated in terms of intelligibility of different length units. In literature, testing of smallest speech units, like phonemes or syllables, is often called *segmental* testing. Tests are performed by playing isolated segments of speech to the listener, whose task is to repeat what he hears, by spoken or written response. Test results are calculated by the basis of amounts of correctly repeated segments. The segments can be meaningful words, or single or set of nonsense phonemes. Tests are sensitive to segmental errors in speech, like false produced phonemes, because there is no context information that would help the listener to guess the segments not heard. Synthesis system development can

then be focused to the segments that performed poorly in tests.

There are no demanding requirements or training needed for the listeners participating to the tests. If something is wrongly heard, it is not because of inexperience of the subject, but because of deficiencies in synthesis. Usually subjects only need to be native speakers of the language tested.

One of the first segmental speech evaluation method was that of *Phonetically Balanced Word Lists (PB)* (Egan 1948) developed in Harvard university during the Second World War. They are lists of meaningful, monosyllabic words with limited range of consonant-vowel transitions. The words have approximately same phonetic content than the language in issue relatively has. PB test was originally planned for comparisons of early telephony devices, but it can also be used for segmental evaluation of speech synthesis. In the test, the word lists are played to the subject, whose task is simply to repeat the word he hears. Devices are compared in percentage of correctly repeated words.

As enhancements to PB tests, the rhyme tests have been introduced. They have became very commonly used segmental methods in synthetic speech evaluation (Lemmetty 1999). In rhyme tests, isolated, but meaningful words are played one by one to the listener, whose task is to select what he had heard from the options in an answer sheet. The answer options are words that are similar to each other, only having a minor difference. Three tests can be found in rhyme test category:

- *Diagnostic Rhyme Test (DRT)* (Fairbanks 1958), in which there are two alternative answers for each word presented. This is the original rhyme test where the answer options are rhymes to each other differing in the initial consonant, like "kill - bill".

- *Modified Rhyme Test (MRT)* (House et al. 1963), lists six alternative answers for each word. An open response sheet can also be used (Logan et al. 1989). MRT is arranged so that in one half of the test the answer alternatives differ in first consonant, similar to DRT, but in the other half the alternatives differ in final consonant, like "dig - din - dip - did - dim - dill".

- *Diagnostic Medial Consonant Test (DMCT)* is similar to DRT, except the answer options differ in consonants in the middle of the word, like "stopper - stocker".

Results of these tests are given as a percentage of correct answers in consonants in issue. Specific results can also be extracted about which consonants and phonetic transitions are inclined to errors.

Tests above are based on meaningful words and need careful preparation of word lists. Every language naturally needs own lists, and it is time-consuming and might be difficult to find appropriate monosyllabic words for rhyme tests. In addition, the result may be

inaccurate if subjects can guess the answer despite of a false perception. These complicacies can be overcome if tests are allowed to include also nonsense words.

Nonsense words in segmental testing are called *logotoms* (Goldstein 1995). They are construed from such consonants (C) and vowels (V), whose synthetic production performance is of interest. Most common logotoms are syllables of form CVC, when testing consonants in initial or final position, and of form VCV, when testing the consonant in middle position. The idea of test list collection is to produce logotoms that follow language rules, but bear no meaning. Logotom structure, frequency spectrum and amplitude distribution should correspond those of natural speech. The lists should be long enough to make them impossible to memorize by hard, and they should be equal in difficulty between each other. Testing with nonsense words can be done by asking the subject to repeat the whole word, or to fill in the missing consonant of answer sheet etc. With this procedure, all the phonemes, and transitions between them, can be effectively tested and erroneous ones can be found. The focus can also be pointed to such special phoneme combinations that are rarely occurring in meaningful words.

Logotoms are also used in *Cluster Identification test (CLID)* (Jekosch 1992). It incorporates a word generator, which produces phoneme clusters from statistical input data. The generator can be adjusted to produce, for example, certain cluster structure (e.g. CVC or VC) with certain frequency of occurrence of phonemes. Because of statistical nature stimulus generation, the generated words mainly are nonsense. The subject hears the words and repeats them in an open response answer sheet. From subjects' answers, the results are extracted for whole words, as well as for initial, medial, and final clusters of phonemes.

As an example of CLID test is that for German synthesizers in (Kraft & Portele 1995). They generated vocabulary of 900 items of phonetic form C^iVC^f, converted items into orthographic presentation, and used those as an input for TTS systems. Achieved stimuli were played one by one to the subjects, who were given the following instruction: "Please write down what you have heard in such a way that another person would read it aloud in the same way as you heard it originally." Since there are no pronunciation rules for nonsense words, some alternative answers, which all could be pronounced similarly, were accepted. The results were presented as recognition rates at word level and at initial, medial, and final cluster levels.

Segmental tests are easy to arrange to focus on certain problematic units. In case of speech processing, consonants are usually hard to handle. That can also be seen in history of segmental test, them being concerned in recognition of consonants, transitions between them, and transitions between consonants and vowels. Segmental tests can predict the entire intelligibility of TTS systems, but do not directly measure the communicative capabilities of those.

3.1.2 Supra-segmental intelligibility

Sentence-level tests evaluate the intelligibility, or comprehension, of longer units than words. Tests consist of lists of sentences, whose content is usually characteristic to the language. The sentences may be meaningful, or they can consist of miscellaneous words producing nonsense entirety. Sentence stimuli are spoken to the listener, whose task is to repeat the whole sentence or only parts of it, for example the last word. The correct perception of each speech unit is not important - segmental tests are used for that - instead the transmission of the whole message is of interest. The listener can miss some parts of the speech, but still give a correct answer with the help of contextual cues.

(Lemmetty 1999) summarizes three different sentence sets that are commonly used in English speech quality evaluation:

- *Harvard Psychoacoustic Sentences* (Egan 1948) is a set of 100 sentences. They are phonetically balanced to match average English and represent typical phrases.
- *Haskins Sentences* are syntactically normal but semantically anomalous (e.g. "the great car met the milk"). If listener fails to hear a word, it is impossible to conclude it from the context. Testing with Haskins sentences is closer to segmental testing than with Harvard sentences.
- *Semantically Unpredictable Sentences* are collected from a list of candidate words with five different rules of grammatical structures (e.g. Subject-verb-adverbial). The sentence lists are thus not fixed like in the cases above. Selection produces semantically anomalous sentences, similar to those of Haskins sentences.

In all sentence lists, the problem is the learning effect. One sentence can be presented to a subject only once during the test, and a subject can participate the tests that use the same lists only once. When sets of sentences are carefully developed, as if Harvard sentences are, they are eagerly used in many types of sentence testing. Ideally, it reduces the workload by letting several tests be run without developing new stimuli, but with time, the same sets become familiar to people in field. For this reason, the participating subjects should be naive.

Listening comprehension is also tested with sentence stimuli, or with longer passages of speech. Nonsense sentences, like Haskins', naturally cannot be used when studying message comprehension. In comprehension tests, the subject hears a speech passage and answers question about the content, instead of repeating what he hears. In this case, the subject needs to memorize or process the message, all of which requires cognition.

The cognitive effort of speech processing can be studied in several ways, for example by measuring the time needed for processing the message, or by inspecting how much

does speech processing reduce some other, simultaneous cognitive task. Speech with better comprehensiveness ought to transfer the message faster and the response to the stimulus should be elicited quicker. However, when asking, whether an individual word occurs in speech passages, subjects' reaction time for synthetic speech may be better than for natural. It is suggested that when listening synthetic speech, more effort is paid to actual words spoken, and ability to interpret and memorize the meaning of message is reduced (Goldstein 1995), so the comprehension task should be more generic. An example of stimulus passage, question about stimulus, and answer options could be (respectively):

"Models '1234', '2345' and '3456' are available",

"Are all models available?",

"Yes / No / Can't tell from information in the sentence".

If performing two simultaneous cognitive tasks, such as speech processing and driving, they will be mutually affected so that higher load on other will reduce the performance in either. This can be used in comparison of different speech synthesis systems comprehension by measuring performances of both tasks.

Naturalness can be defined as how much the synthetic speech is similar to the natural speech. Naturalness as well as overall quality of synthetic speech are abstract subjective attributes, which are not easy to quantify, because subjects most likely find different aspects of either more important. As discussed in the beginning of this chapter, each system under comparison can have the best quality in certain perspective. Naturalness can suffer from several deficiencies in speech and systems may sound natural in different manner. In TTS systems, intelligibility and naturalness are highly correlated (Klatt 1987), so good performance in intelligibility tests may suggest good naturalness, although it is possible that a system sounds natural, perhaps because of good prosody, but is inferior in segmental intelligibility.

For evaluating the subjective attributes, a paired comparison can be arranged to obtain judgements of preference. This was done by (Kraft & Portele 1995), who performed a pair comparison and ranking of synthesis systems. Subjects listened to the systems pair wise, and selected the one they preferred regarding to the intelligibility combined with easiness of long-time listening. The outcome was that intra-subject results were consistent, but inter-subject results were not, because too many degrees of freedom were left to judge.

Superior method for evaluating the subjective attributes, such as naturalness, is mean opinion scoring. Because of its importance in TTS quality evaluations, it is discussed in more detail in section 3.2.

The suitability for particular application can be concluded by appraising the application needs. They are often more pragmatic than those of interest in the research, such as the storage and computational requirements. Unit selection with large acoustic inventory is

problematic in hand-held devices, but not an issue in industrial server frameworks. The same concerns the computation needed for articulatory synthesis. There have also been discussions if the naturalness of synthetic speech is desirable. It might be more comfortable for humans to recognize easily that they are listening to, and speaking with a machine, instead of mistakenly believing that it is another human, due to very natural synthesis.

3.2 MOS

An easy way to evaluate sound quality is to let a person grade some characteristic of sound, for example by asking, "What grade would you give for this sound quality?" To get somehow reasonable answer to this question, it needs definition that is more exact. The scale of grades has to be known as well as references about what is good and what is bad in which sense. The references could be the extremes of the scale of predefined aspect, and the improved version of question becomes "In scale from 1 to 5, the first sound is of quality grade 1, and the second sound is of grade 5. What grade would you give for the third sound?" The reference sounds are called *the anchors* and they lock the known levels. Besides the extremes, another way to set an anchor is to play an average quality sound and set it in the middle of the scale. Then the problem is that the respondent does not know the extremes, and the extreme scores might not be given, because it cannot be predicted if some other sound is even better or worse. However, if there are several sounds under judgement, the entirety of available qualities will be found eventually, and whole scale will then be effectively used. The best, but also most time-consuming way to set the anchors is to play forehand a variety of all sound qualities to the respondent, as many times as it is necessary for him to understand the scale. Only after that, the question is pointed to the sound to be judged.

MOS (Mean opinion score/scale) test, a very widely used evaluation method in telecommunications, is based on the grading procedure. Several subjects will grade sounds in comparison with the anchors, and an average of given scores is calculated. The average represents the common opinion of quality. Instead of just one overall quality grading, the test is usually formed as a questionnaire presenting questions about several characteristic items of sound, such as its pleasantness, clarity, or overall impression. Subjects concentrate in one item at time and grade it when the sound sample is played. The sample may be continuous, so that the subject may answer the questions at own speed during the playback, or there can be short breaks between samples, while subject answers the question and reads the next one. The latter way is prevailing.

A typical scale in MOS is 5-point wide. In the answer options in questionnaire presented to subject, the numerical values from 1 to 5 are replaced with respective written descriptions

from "excellent" to "bad", or similar, depending on what is asked. The questions may be pointed directly to the quality, or to the degradation of the quality, whereby the best grade is "no degradation" and the worst is "very annoying degradation". In this way, the quantitative results can be found with plain qualitative testing procedure. Quality descriptions, which are connected to the quantitative values, are called *phrase anchors*. In many cases, the term anchor matches the level-locking definition above. For example, when asking "Did you hear any abnormalities in speech?", the scale can be from "no" to "yes, they were very annoying" That scale is locked: the subject might not hear abnormalities, or he might hear them, even very annoying. In some cases however, the definition might fail, such when asking "What is the overall quality?" with scale from "excellent" to "bad". First test sound might be "excellent" compared to second, which in turn is "excellent" compared to third. Being aware of this, an auditory anchoring is needed in addition to phrase anchors.

The test performance time becomes longer when evaluating several systems in scope of several sound characteristic items, instead of just one overall item. Total amount of questions needed is *amount of systems* $*$ *amount of items*, so that every system will receive each of the questions. To get the respondents accustomed to the quality variations between systems, all the audio samples may be presented in such a pseudo-random order, that each question is pointed once to each system.

Benefit of MOS test is its efficiency and that it is easy to perform. A group of people can simultaneously attend to test, so a sample of results can be achieved already by running test once. Unless there are specific needs for accurate anchoring forehand, the test can be performed without lots of time-consuming calibration and guiding. The respondents need not to be specialists, since their task is to represent average people and give their subjective opinions.

In telecommunications, MOS test has traditionally been used for quality evaluation of speech coding algorithms and transmission channel interferences. In those, the quality is depending on the system, and MOS test is used to distinguish between them. If other systems receive better score than others, they are generally appraised to have better quality. Being an efficient subjective testing method, MOS is an obvious choice for auditory quality evaluation of TTS systems, also recommended by International Telecommunication Union (ITU 1994). In the recommendation, the items related to the sound quality that are to be scored are:

- Global impression
- Listening effort
- Comprehension problems
- Speech sound articulation
- Pronunciation
- Speaking rate
- Voice pleasantness

Each item measures the quality of corresponding character in TTS. Simple average of all items gives a single score for comparisons, but also a prominent combination of several items can reveal important aspects of the system. These combinations usually appear to be mutually dependent, so if one item receives good scores, also other in the same group will. There has been discussion about how the items can be combined and what do the combination scores actually stand for.

(Kraft & Portele 1995) performed tests for five German speech synthesis systems. Their category-rating test was following the MOS procedure with eight items, that were "pronunciation", "comprehensibility", "distinctness", "intelligibility", "speed", "pleasantness", "stress", and "naturalness", as translated to English. For anchoring, two sentences from each system were played before the actual test. The test material was selected to be passages of speech, each of about 100 words long and equal in difficulty of comprehension. For one system, all the eight items were judged during two passages speech, from which the other passage was common to all systems. The performance of the test was quick: 13 minutes / individual subject.

After running the test, the results were statistically analyzed in aim to find the underlying main factors of the sound quality and in that sense reduce the amount of items. The idea was to find a smallest possible set of features that will capture all the important differences between the voices. Eight tested items were combined into two, "segmental" and "prosodic" with help of statistical factor analysis. From eight items presented above, the segmental group included the first four, and the prosodic group included the three last. "Speed" did not belong to either group, thus suggesting that speed is not affecting segmental or prosodic features of speech. Statistical analysis however showed that speed is more towards segmental group.

The study of (Polkosky & Lewis 2003) aimed to revise the ITU recommendation of MOS extent. Taking also into account the factor categorization of (Kraft & Portele 1995) study, they prepared two new scales, MOS-R (Revised) and MOS-X (eXpanded), from which the former was an intermediate phase to the latter. Their procedure was to run the MOS test, search for underlying speech quality factors by the basis of the results and add new items to the scales, until a sufficient correspondence between the items and the factors was achieved.

The final MOS-X test they suggested has the items grouped following way:

- **Intelligibility:** Listening effort, Comprehension problem, Speech sound articulation, Precision
- **Naturalness:** Voice pleasantness, Voice naturalness, Humanlike voice, Voice quality
- **Prosody:** Emphasis, Rhythm, Intonation
- **Social impression:** Trust, Confidence, Enthusiasm, Persuasiveness

(Viswanathan & Viswanathan 2005) criticized the studies above with reasonable arguments. They presented a good review of test generation procedures in *psychometrics*, which is the use of statistical analysis to evaluate the quality of practitioner-created scales and other psychological measures (Polkosky & Lewis 2003). The fundamental error in the MOS studies, where factors are extracted from the results of single items, is that there have been many measurements before knowing what is actually measured. (Viswanathan & Viswanathan 2005) resulted to a state-of-art MOS test for TTS systems. In their work, the path between an abstract concept and its concrete measurement was carefully traversed. Moreover, the exploratory factor analysis used to factorize the items in entirety in previous studies, was extended to exploratory factor analysis of preliminary stages of test development, and to confirmatory factor analysis for confirming the results belonging to the predefined factors. MOS test is suggested to have only one factor or two, if a measure that is more accurate is needed. The two-factor model, and items belonging to those, has minor changes to earlier studies (see Appendix A for the complete subject questionnaire):

- **Naturalness:** Naturalness, Ease of listening, Pleasantness, Audio flow
- **Intelligibility:** Listening effort, Pronunciation, Comprehension, Articulation, Speaking rate

3.3 SRT

A traditional need for speech intelligibility measurements is to test people's degree of hearing impairment. The point of interest is not the signal quality in transmission channel or the quality of the speech source, but the receiving end's ability to understand the delivered message. A challenge in such measurements is to develop a test that describes actual communicative capabilities of the listeners, instead of, for example, just trying to find out if the listener can satisfyingly hear tones in the normal frequency scale. The objective of a test should be to find out how the listener can understand speech in every day situations. A natural way to do this is to make the subject listen speech and somehow estimate how

well he understands it. It is also clear that tests where a series of individual, unknown or unpredictable words are presented to the subject, and the subject is asked to repeat what he hears, don't represent the objective above. Instead, the test corpus should consist of longer units, such as sentences and small passages of general speech that are logical and characteristic to the language. This is more close to the every day situations, where the message can be understood by the listener although he might not hear every word of it correctly but can conclude the missed words with help of the cues in the context.

Natural hearing event is always affected by several different interferences in transmission channel. Even a simple case when a human is talking to another in a silent room, there is most likely some ventilation humming etc. pushing in. An extreme case would be a conversation in a cellular phone beside a road with lots of traffic. In that case the transmission channel has a huge portion of noise: distortions in conversion between an acoustical waveform and a transmissible data flow, noise in telecommunication channel and acoustically coupled noise in both ends. This makes the speech very hard to understand and suggests that the speech intelligibility measurement could be done by adding noise to the speech stimulus in some controllable manner.

(Kalikow et al. 1977) developed a test to measure person's ability to understand speech, especially of noise disturbed. The test consisted of sentence material to be presented to the subject simultaneously with masking noise. One of the first questions in their task concerned the type of response being elicited from the subject. The response had to reveal the intelligibility of the stimulus as what the subject understands. They also wanted to keep the subject's task simple and asked only to repeat the last word in every sentence presented. The word at issue was restricted to be a monosyllabic noun, so that the main stress of the sentence was received by it. They used two types of sentences: highly predictable (PH) sentences, such as "the beer drinkers raised their mugs", in which the last word can be guessed from the remainder of the sentence to be something like "glasses", "jars" or "mugs", and low predictability (PL) sentences, such as "I should have considered the map", in which the last word can basically be whatever. This division was aimed to distinguish between the subject's cognitive processes and sheer acoustic-phonetic information, which is not dependent of thinking the answer so much. The phonetic profiles of the sentences were analyzed to be representative of English. Interfering noise was selected to be a babble of 12 speakers reading continuous text. It was found to confuse the speech more than stationary random non-speech noise, since it gives false speech cues and adds the load on the attention and memory processes.

As an outcome of the work, the SPIN (speech perception in noise) test was introduced. It consists of a set of test sentence lists and a two-track recording with the sentences spoken on the other track and the babble noise on the other. The test is to be performed in a fixed

SNR, so that all the recordings hold the same level, but the noise can be changed from test to test. The results of tests are presented as a percentage of correctly repeated last words for both PH and PL sentences in SNR used. The combination of these scores has the potential of predicting the ability of the hearing-impaired to perform in every day situations. The major advantage of the test is the rapidity: one form of 50 sentences in one SNR can be administrated in about 10 minutes. It was also suggested that the small difference in PH and PL scores might reveal the possible defeat in cognitive and memory processes. This means that the subject cannot perform any better with the sentences where he can at least guess the last word compared to the sentences where the word has really to be heard. However, these suggestions have not received support, since they do not seem quite reliable quantity them selves.

The fundamental drawbacks of the SPIN method are the floor and ceiling effects of the results. In a correctly formatted test, the intelligibility of the stimulus should be a linear function of SNR, that is, the better SNR, the better result in relative amount. However, as the sample result of the SPIN test in Figure 3.1 shows, above and below the certain level of SNR relatively more or less words are understood, respectively.

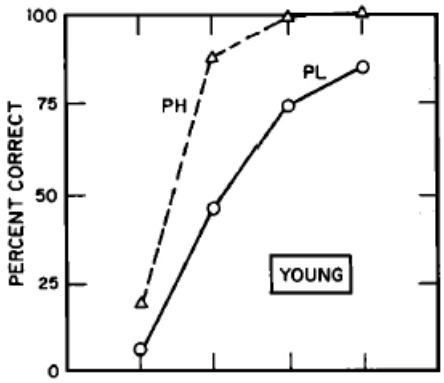

Figure 3.1: Example of SPIN test results.

An alternative measure to percent intelligibility is SRT (Speech Reception Threshold), which also measures the listener's ability to understand speech. The name is descriptive: the aim is to find that speech volume level that is barely intelligible. Subject's SRT can be found with an adaptive test where the signal presentation level is adjusted while a sequence of speech stimulus is played to the subject and the subject is repeating what he hears. If the answer is correct, the next item in the sequence is played at lower volume and if the answer is wrong, the next item is played at higher volume. This procedure converges towards the threshold level where the subject just hears the speech, namely SRT. At the era of SPIN test, also SRT test was introduced by (Plomp & Mimpen 1979). In their test the speech stimuli

to be repeated by the subject were whole sentences, and therefore the method has also been known as sSRT(sentence speech reception threshold). They prepared and recorded 10 Dutch list each containing 13 sentences that are to represent conversational speech, are short enough to be easy to repeat, and are neither too redundant (e.g. proverbs) nor too difficult or confusing. All the lists were tried to be arranged so that each list has equal number of the various phonemes. To get the test useful for various conditions, the frequency spectra of the sentences were analyzed and noise having the same frequency content was recorded.

The procedure for measuring the SRT is following. At the beginning, the first sentence of a list is played repeatedly with increasing sound level until the subject repeats the sentence correctly. The sound level is decreased by 2 dB and the second sentence is played. From this on till the end of the list, the level is decreased or increased by 2 dB regarding if listener can repeat the sentence correctly or not, respectively. The average level of sentences 5-14 (the item 14 is not present, but the level is known from the previous item) is calculated and that is the SRT for the list. The overall SRT can be found by averaging several different list results. Naturally, the same list can be used only once for a subject.

The advantage of the SRT test is that it is not suffering from the floor and ceiling effects like SPIN, as reported by (Nilsson et al. 1994). In comparison with SPIN test, SRT test has more accurate linear correlation between SNR and speech intelligibility. While in a percentage test of SPIN type almost all stimuli are understood after increasing the SNR over certain level, the SRT test does not spent so much time moving in this non-linear area.

The benefits of SRT test has led to further interest and research of the issue. A common development area is the sentence lists. Proposed lists may need enhancements or updating. There have also been discussions about how many sentences should each list contain and how the SRT is then calculated from the answers. Different languages naturally need own lists and those have to be collected, recorded, and processed following the same or improved new standards. Since this is somewhat a difficult and time consuming task, some attempts has been made to reuse old test material, such as SPIN sentences, as a SRT test corpus (see (Dubno et al. 1984) and (Gelfand et al. 1988)). Finding the limitations of these approaches, (Nilsson et al. 1994) developed a set of American English sentence materials specifically for use in SRT measurements. They started to collect the material from Bamford-Kowal-Bench (BKB) sentences, which are a large set of short sentences incorporating common nouns and verbs in British children's speech. They resulted 25 equivalent lists of ten sentences that have been normed for naturalness, difficulty, and reliability, and entitled this test as Hearing In Noise Test (HINT). The sentences were spoken by a male professional voice actor and recorded. The masking noise used in SRT task was chosen to match the long time average spectrum of the recorded speech, since it approximates the babble noise and ensures that on average the SNR ratio will be equal at all frequencies.

(Vainio et al. 1994) pointed criticism into HINT test sentence lists, mainly because BKB based sentences do not represent current adult language usage, although the adults are most likely target group. Accepting the methodology itself, they prepared new HINT lists in English and Finnish. The aim for the work was to develop 25 lists of 16 sentences in each language. The source material for both languages consisted of sentences containing several million words collected from newspapers and other sources that use current adult language. All the materials were converted into phonetical form, which is easy for Finnish, since the correspondence between the spoken and the written language is high. For English, the Festival speech synthesis system (*Festival home page* 2006) front end was used. Omitting too long or short sentences, a few tens of thousands from the whole material were left to be parsed more carefully. Grammatically incorrect sentences were removed. Biphone (combination of two phones, not to be confused with the term "diphone" used in 2.2.1) frequencies of the remaining sentences were compared to the corresponding of the whole source material. Sentences that had very rare biphones or had the biphone occurrences deviating highly from the average were removed. Similarly, the sentences that had enormous amount of base forms of words were removed. Finnish sentences containing words with unpredictable pronouncing were removed and English sentences were checked for naturalness. Per language these filtering left approx. 1000 candidates that were used to achieve the final 25 lists having 16 sentences each. At first stage, the lists were created randomized, and then the sentences were started to swap pair wise so that each list was balanced for average sentence length, average word base form frequency and distribution of phonemes. All the changes that reduced the variance of the variables significantly were kept, until no improvement was observed.

The lists were recorded by four talkers: Finnish and English male and female. The average volume levels of the recorded sentences were equated. The masker noise long term spectrum was calculated for both languages by summing up all the sentences of the language at issue. Infinite impulse response filters were designed to match 128 frequency points of calculated spectra and those were used to filter white noise.

Despite adjusting the average volume to be the same in all sentences, it was expected that the overall intelligibility of the individual sentences would not be equal. Therefore the sentences were subjectively tested for intelligibility and scaled up or down a few dB according to the results. An SRT task was arranged for the final lists. It was proved that all the lists gave equal result as an average of the levels of the sentences 5-17 (again, item #17's level is known from the previous one). In this phase, all the lists' SRT should have been equal, but still some variations were found. It was thought to be due to linguistic differences and differences in listeners and the talkers who were speaking the sentences. This was however overcome by omitting two lists in Finnish (lists 15 and 16) and one list

in English. Thus the ultimate result of the work was sentences lists of 16 items, 23 lists in Finnish and 24 lists in English. It was claimed that a suitable list length could safely be reduced from 16 to 10 items.

Generally in research, the measured SRT value tends to be a few decibels below zero. This means that the speech is such a strong stimulus to human brains, that especially when concentrated on, it can easily be reconstructed from noise that is a lot more powerful than the speech. The reliability of a SRT test can be indicated as the variation of several lists SRTs. Perfect test would give the same SRT value for all lists, so that the variation is zero, but in practice this is not the case. The standard deviation of different lists SRTs in tests described above is around 1 dB.

3.4 Scope of this thesis

The aim of this study is to compare the selected Finnish tts-systems in terms of HINT test described in Sec. 3.3. However, parting from original HINT, this time the quantity to measure is not the subjects' ability to hear and understand the stimuli presented. In fact, to some point, it is assumed that all the test subjects would have normal hearing and thus very similar results. The point of interest is the speech production of the synthesizers. Test composition will follow closely that of HINT, except that the sentence lists are not spoken by a human. Instead, they are fed to the tts-systems, which produce the recordings. If the SRT's of different synthesizers vary considerably regardless of the listener, it can be assumed that better performing synthesizers are more intelligible than others.

The use of HINT in the evaluation of speech synthesis is a new approach to the issue. The focus of synthetic speech intelligibility testing has usually been in segmental evaluation and in MOS testing. Therefore, the applicability of HINT in synthetic speech intelligibility evaluation will also be studied in this work.

An interesting question arises about the difference between linguistics-to-speech conversion techniques. The preconception is that unit-selection type synthesizer (see Sec. 2.2.1) produces very monotonic speech lacking most of the prosody. This causes the synthetic speech effectively vanishing under the masker noise. The speech with proper prosody modelling will instead give the listener better prosodic cues in sentence level and is therefore more intelligible under noise.

Another question is that can a synthetic speech in some sense have better intelligibility than natural speech. At prosody level, this is hardly possible at time. There are deficiencies in models trying to copy the natural syntactic prosody, and semantics are not modelled at all. In segmental level, however, there are well-defined theories about what acoustic properties correspond to human speech perception. For example, the formants and their clear

distinction in speech signal are important. If those can be emphasized in natural speech, there is a possibility to achieve synthetic speech that is superior in intelligibility. Concatenation of natural speech segments has the formant structure of the speech underlying it, and definitely cannot take any enhancing advantage from mutilation. Apposed to that, the synthesis by rule is free to do whatever while creating sounds, and can be set to emphasize the formants in resulting speech. (Klatt 1987) claims that all trials to find synthetic "super speech" have failed, because although some cues in speech are more powerful, the listeners appear to be responsive to acoustic details, which are known regularities.

In addition to HINT, also a non-formal MOS test was prepared with a few participants. HINT is only intended to discriminate the systems by the means of their intelligibility, but also other aspects of speech synthesis quality were of interest. The MOS followed closely the test suggested by (Viswanathan & Viswanathan 2005).

Chapter 4

Methods used in this work

4.1 Fetching material

The aim of this works is to compare Finnish text-to-speech systems and find out their performance in intelligibility in terms of HINT test. This chapter describes the procedures used to select the materials for the study. It includes the TTS systems, the text material for the systems to synthesize and the masker noise used in HINT.

4.1.1 TTS-systems

First requirement for the speech synthesizer systems was their ability to read plain text input correctly. This rule limits out the systems in development state, which need extra information, such as control tags among text, to produce the speech. Definition of this was easy: feed the text file in and get the speech file out. Secondly, the participating synthesizers must be generally available.

Markets for the Finnish speech synthesizers are small and therefore there are not so many competitors in the area. Actually, there were fewer of them than desired so that better selection criteria could have been used. Current situation yielded four different systems, three concatenation based, and one of parametric type. An interesting comparison would have been between several different types of low-level synthesis methods.

All companies that provided TTS systems into test were instructed to supply the best possible quality systems available. So that the rules could be followed strictly, it was preferred that administering personnel could have an access to all systems. In practice, it means that all systems were installed to a pc.

1. **Infovox**, henceforth referenced as "tts1", is a Swedish company, whose product "Infovox Desktop" can speak Finnish. It is a diphone concatenation system based on

MBROLA (Multi Band Resynthesis Overlap Add), which is a version of PSOLA described in Section 2.2.1.

2. **Mikropuhe**, "tts2" is a speech engine, which is used as a computer desktop reader product and, for example, as public entertainment in form of a "talking head" seen in TV. It is based on concatenation of "microphonemes", which are recorded speech segments of length about 10 ms (Lemmetty 1999). Concatenation is pitch-synchronous of some kind, because the speaking rate and pitch can be freely adjusted. The pitch adjustment property also offers a possibility to make this TTS system sing by adding control tags of musical notes to the input text.

3. **Bitlips**, "tts3" is the only parametric TTS system of the group. It is based on HTS synthesis, so that it produces speech from mel-cepstral coefficients and excitation parameters emitted from concatenated HMMs (see Section 2.2.2).

4. **Puh.e**, "tts4" is a Finnish extension to the product "IBM Websphere Voice" based on "the IBM trainable speech synthesis system" (Donovan & Eide 1998). It has the roots in the work described in (Donovan 1996). The system uses context-dependent unit selection from a speech database prepared with help of HMMs.

Characteristics of each system can be examined from the Figs. 4.1, which represent long-time spectra, and Fig. 4.2, which represents F0 contour estimate and intensity contour. The figures are extracted from each system speaking the sentence: "Pohjantuuli ja aurinko väittelivät kummalla olisi enemmän voimaa." Some of the properties of the systems can immediately be seen. The formants of the speech can approximately be studied from the spectral pictures. Long-time spectrum for tts3 shows highly emphasized resonances at frequencies about 2.5 and 3 kHz, which are common frequencies for F3 and F4. These peaks occur at band that is sensitive in human hearing, and thus the system is likely to be more intelligible than the others are. In the systems, which produce speech from mel-cepstral coefficients, the formant structure of the speech can be emphasized by adjusting the expansion parameters. The expansion may partly explain the resonances, but because the natural movement of the formants would eventually produce flat spectrum, there might be other kind of filtering present in tts3. Tts4 especially has very flat, descending spectrum similar to natural speech. The spectrum includes also lots of components in the high frequencies, most likely because of the distortion in concatenation joints of the units.

The intonation of speech can be seen in estimates of F0 contour. Tts1 has unnatural intonation, being very extensive, simple saw-tooth shaped and rapid. Tts2 has similar shape, but in more calm manner, and also incorporating descending trend. Tts3 and tts4 both have finely built shape, even slightly smoother for the former, but also a good achievement for

the unit-selection. In intensity, the unit-selection is more typical: the speech is more like on-off type lacking the variation. Worth of noticing is that tts3 has mutually correlated F0 and intensity contours: the emphasis occurs simultaneously in stress and intonation. Others have curves that are mutually more independent that is less natural behaviour.

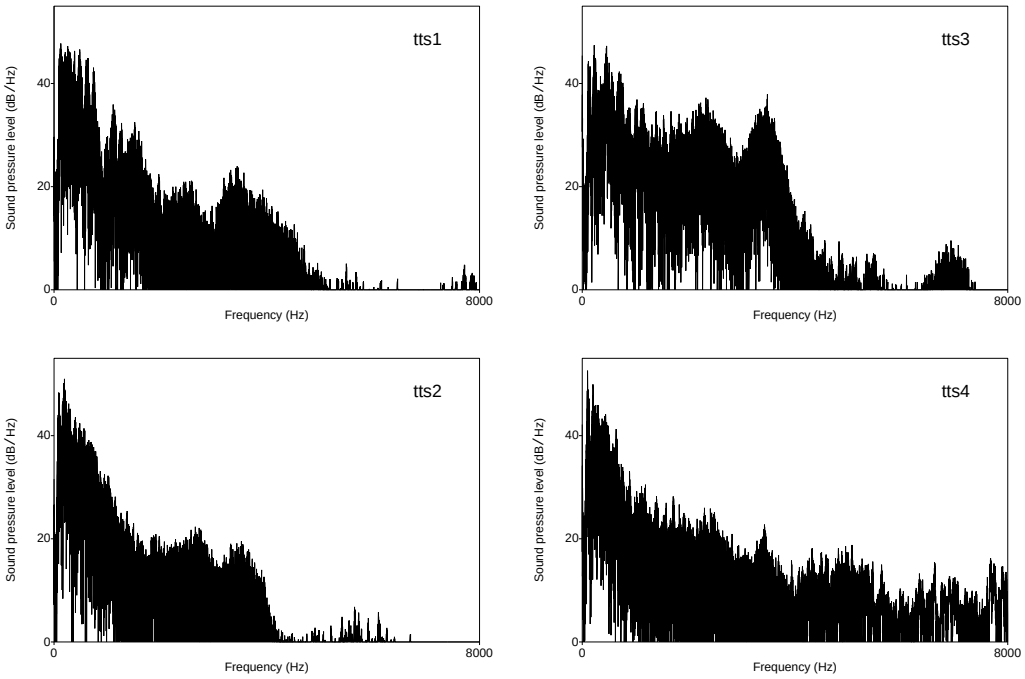

Figure 4.1: Long-time spectrum of each system speaking the same sentence.

4.1.2 Text material

Text material selected to the tests were adopted from (Vainio et al. 1994) research as described in section 3.3. They had prepared 25 sentence lists all containing 16 sentences. The performed test is based on those lists, which are this time read by the TTS machines, instead of the humans as in the work in citation.

Selection of amount of text material had to be made as a compromise between reliability of results and reasonable level of test subjects' exertion. Presenting only one list per machine to a subject gives an SRTvalue, but it probably is inaccurate due to small sample taking. In addition, it was pointed out in the original study that there is intelligibility variation between the lists, because of unknown reason. Therefore, it was preferable to present as many list as possible to find the average SRT with small variation for each machine. The maximum amount of lists to present comes from the fact that subjects' available time is

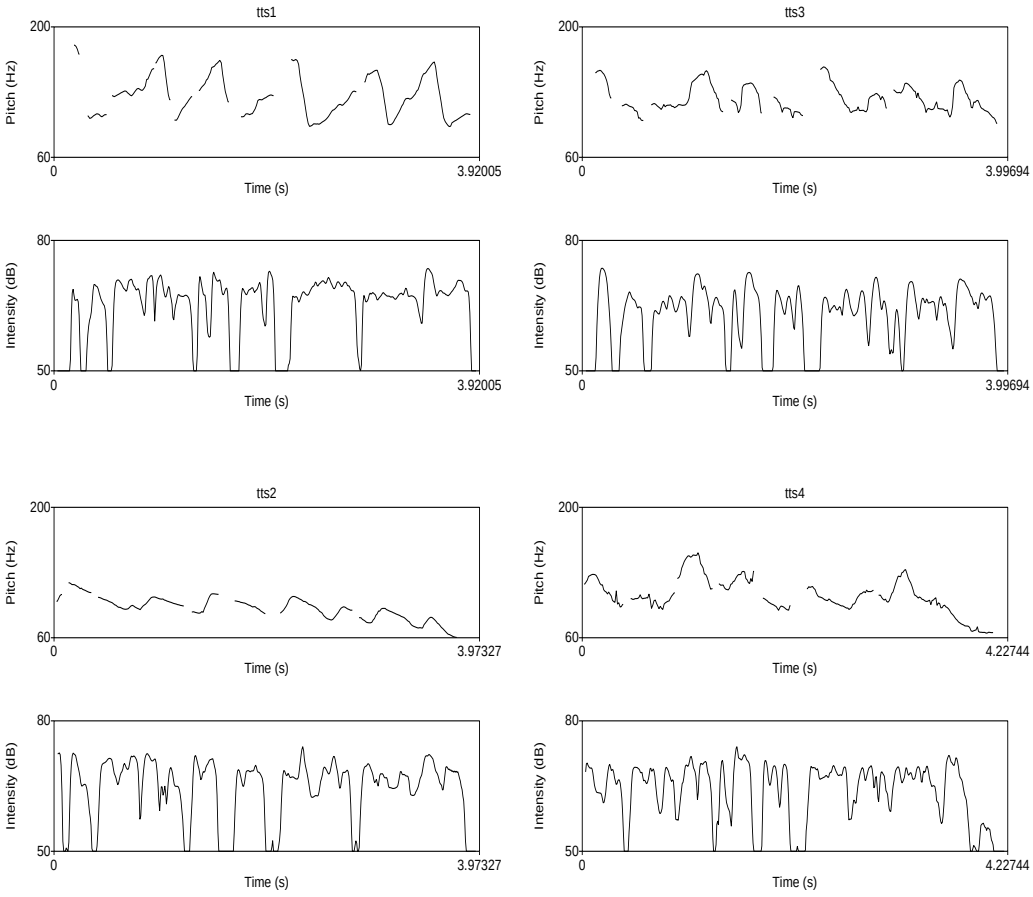

Figure 4.2: F0 contour estimate (upper figure in a pair) and intencity contour (lower figure) for each system reading the same sentence.

limited: they do not want to spend too much time and the answers might suffer if the subject gets tired or bored. At the beginning, it was decided that each machine should handle four lists, so the total amount of lists presented to a subject was 16. Estimation of the test duration for a subject was one hour.

It was not clear if different lists will have different SRT-values. A solution to this was to present the same lists to all subjects. This way it can be found out if a subject or a list differs noticeable from others. Since all lists should be equal, the first available lists were selected. In addition, each machine got one spare list, in case if subject fails one of the planned lists. The division formed as follow: lists 1-5 tts1, 6-10 tts2, 11-15 tts3, 16-20 tts4. Lists 15 and 16 are reported suspicious in (Vainio et al. 1994), so they were left as spare lists for respective machines. List 1 was noticed to start with a nonsense sentence: "Ohi, ei silläkään ollut asiaa", which was not found to be reasonable in any context. Therefore, that

list was left as a spare for machine 1. Tts4 spare was list 10. The complete lists used can be found in Appendix B.

4.1.3 Noise

Tests for evaluating speech intelligibility in noise have traditionally used multi-talker babble. It is advantageous, since it produces false speech cues and effectively interferes with the speech under test. Creating a babble noise record is easy with even simpler equipment, by simultaneous narration of several people or by using a mix of several voice recordings. The drawback of babble noise is its frequency content being unpredictable. It can happen that some speech is more distinctive among babble, than other is. For example, if there is a babble of several male speaking, it is easy to understand that an additional female voice will be more distinguishable than additional male voice.

The HINT test uses spectrally shaped masker noise. It ensures that on average, the S/N ratio will be equal in all frequencies. The noise has to be created individually to every voice. It is created by first determining the long-term spectrum of the recorded sentences and then filtering white noise with a filter that is calculated following corresponding smoothed spectral curve. Spectrally shaped masker noise has such an advantage over babble noise that it is tailored to each voice, and none can take a benefit of it.

Another noise type, which is even to all voices, is white noise, and it was used in this study. White noise is such a statistical signal, where the expected value for each frequency component is equal. Thus, the spectrum of the white noise eventually approaches flat. The average power of noise is dependent on the expected amplitude values and the frequency band the noise occurs or is measured. To get the white noise measurable and comparable to the speech for this study, it had to be band-limited. Speech signal consists of significant frequencies only below a few thousand hertz, except only some unvoiced noise-like consonants that can have broader frequency content. Therefore there was no reason to have very broadband white noise for speech masking. The noise selected to the study was computer produced pseudo random white noise filtered to contain the frequency band 0 - 8000 Hz. The frequency spectrum of the noise is presented in Fig4.3

White noise introduces a constant noise floor to the band it is affecting. The volume of the noise can be adjusted for different floor levels, and decreasing it to minimal equals the silence. Speech intelligibility assessments have also been performed in silence, or without mask (Kalikow et al. 1977). It also requires presentation of speech stimuli in minimal sound level. In this study however, a conversational level of the speech stimuli was desired. In addition, it was doubt that scaling very low-level digital signals would be inaccurate, because the quantization error sizes become comparable to the signal levels.

Selecting white noise as the masker was suitable for this work. The aim is to compare

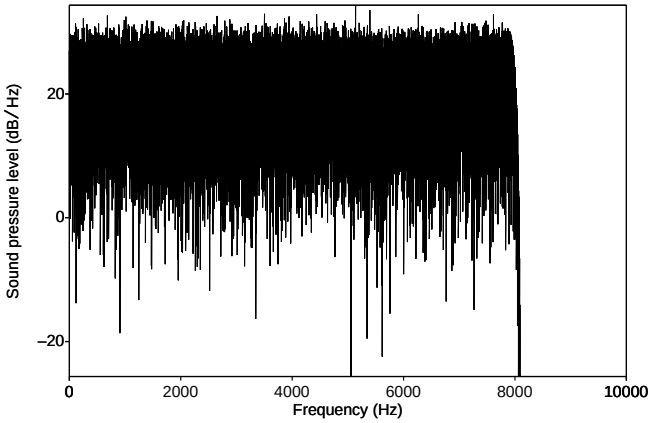

Figure 4.3: Filtered white noise used as a masker in this study

TTS systems' performance in noise instead of studying human hearing. The accuracy of the results does neither need to reveal any hearing impairments of the listeners, nor to be comparable to such research, but only to rank the order of the systems. White noise is also a common interference in communication channels and thus kind of an environment where TTS systems are likely to be used.

4.2 Preparing material

All the audio material produced by the synthesizers had to be further processed to get their properties comparable to each other. The speed and the volume of the speech were different in each machine, as well as the data resolutions in the output files they produce. It appeared that the sampling rates of different synthesizers vary between 16000 kHz, 22500 kHz and 44100 kHz all having the quantization of 16 bits.

Adjustment of properties of speech is somewhat easier for synthesizers than for natural speech. The machine can read text as long as needed with the same characteristics of voice without getting tired etc. It can also be set to recall and repeat settings later on. Only well-trained human speaker can regulate the speech he produces in some degree of accuracy, but still it will have at least minor alterations. It is necessary for the nature of the SRT test to get at least the volumes of all machines equal, but it also nice to have the speaking rates similar for each speaker. Unfortunately due to having concatenative synthesizers involved, it was not desirable to adjust the speaking rate much, because it would make them use additional signal processing that might decrease the quality. Moreover, it was also assumed that a commercially published product should have the best rate set as a preference.

As a common reference, a Finnish fairy tale "Pohjantuuli ja aurinko" of length of 15 short sentences, was used. All the synthesizers read the tale into audio files, and the properties of the synthesizers were analyzed by the basis of the files.

The lengths of the spoken tales were 35-38 s, except for tts2, which was a much slower. However, after adjusting the speaking rate roughly faster, the speech got more pleasant to listen intuitively, and the tale length decreased to 38 s. The speaking rates were reviewed to be similar enough with yet a few individual sentences.

The test software is playing a sentence and the masking noise simultaneously. Different sampling rates of the synthesizers needs to be considered, since it is not possible to play two files having different rates at the same time. Either the masking noises have to be designed separately for all the synthesizers, or the files produced by the synthesizers have to be equated to the same rates. Latter approach was selected so that all the materials were processed to have equal sample rate of 48 kHz. The noise was originally designed as so, but all the sentences had to be up sampled, or interpolated, to meet the requirements. This was done with help of Linux Sound eXhange (SoX) utility. It's "resample" function utilizes the technique called "bandlimited interpolation" as described in (Smith 2002). Upsampling has the benefit of not attenuating any frequencies of the original signal, so despite the sampling rate conversions, the systems' original sound properties were maintained.

There are several approaches to the speech volume balancing. Traditionally in sSRT tests, the sentences are first recorded and scaled equal in mean volume. Then the sentences are initially tested for intelligibility and the ones, which are found most easy to understand, are scaled down and vice versa. After this, the sentences have different total sound power, but they should be equally difficult to understand. It is not clear what creates these variations in intelligibility between the sentences. It can arise because the listeners understand certain sentences easier, or because of the voice quality of the speakers, who were used in the recordings, may vary along the lists. Some sentences content may be more common to an average listener. In a way the message in those are more reasonable as a standalone, while other sentences would need a correct context environment, as they are more unpredictable alone. One reason for the intelligibility variation of the sentences could be the inaccuracy in the volume scaling. Some sentences might include greater amount of continuous vowels that makes them have portions of constant floor power. Some sentences instead, have many unvoiced stop consonants and silent gaps between them, and therefore smaller mean power, although both sentences, the vowel-oriented and consonant-oriented are subjectively equally loud. Just equating the average amplitudes results in having the words sound subjectively unequal. As an example of this behaviour, in Figure (4.4) there are time domain presentations of two Finnish words, spoken by a synthesizer with the same volume setting. The first word, "ääliömäisyys", incorporates several adjacent vowels and thus lots of power.

The second word, "nakkimakkara", has several stop consonants and due to the silent parts, its average volume is lower. Despite of removing all the silent gaps from the latter spoken word, the calculated mean amplitude is 10 % bigger in the former word, although human ear tends to hear them equally loud. However, the sentence lists are phonetically balanced so that this kind of error should be minimal.

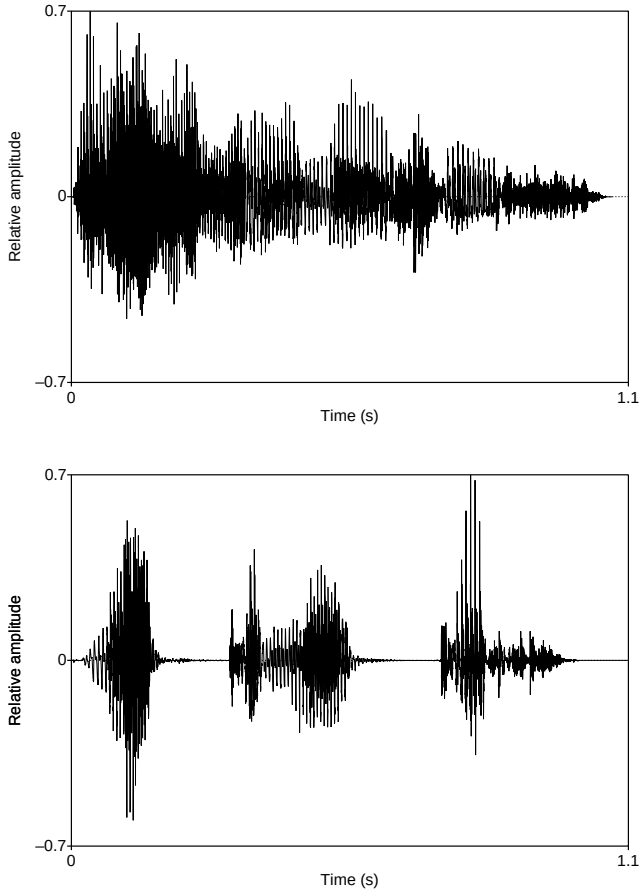

Figure 4.4: Time domain presentation of two Finnish words, "ääliömäisyys" (upper) and "nakkimakkara" (lower)

Since there is no clear reason why the human-recorded SRT test sentences vary in intelligibility, the synthesized sound files were decided to be scaled equal in mean amplitude. Other option could have been to scale the sentences with the coefficients of human recordings, but it might accumulate the speaker specific error to the results. Quantization of the files is 16 bit, which means that there are $2^{16} = 65536$ possible signal levels, so the largest possible signal amplitude is 65536 times larger than the smallest possible. In terms of decibels the difference between the largest and the smallest amplitude (or alternatively its

square, power) level is

$$L = 20\log(\frac{A_1}{A_0}) = 10\log(\frac{A_1^2}{A_0^2}) = 20\log(65536) \approx 96dB \quad (4.1)$$

The limits of signal amplitude are quantization error at the lowest levels and clipping at the highest levels. The average of the speech files was chosen to be 60 decibels above the minimum level, since then it is possible to increase and decrease the volume without running into limits immediately. In dB representation, the same level quantity can be regarded to be either amplitude or power level, as can be seen from Eq. 4.1.

From this point, the Snack Sound Toolkit (*Snack home page* 2006), and a tcl-script "normIt" (Rohde 2006), which takes advantage of Snack, were used to perform further processing of the audio files.

Tts2 appeared to produce two-channel files instead of others' mono files. Snack has a channel conversion function which in this case simply calculates an average of both channels' samples to produce a new, one channel file: $x_n = \frac{x_l + x_r}{2}$.

Because the sentences might have linguistic differences that reflect into power differences as described above, it was preferred to make the synthesizers speak equally loud on average, instead of scaling individual sentences to the same level. The average power of the signal can be calculated by summing and squaring the samples: $L_{MSx} = \frac{\sum x_n^2}{n}$. The averaging was done by letting the synthesizers first speak all the material at their default constant volume. The out-coming sound files were scaled with a machine-specific attenuation factor, which was calculated from the reference speech, "Pohjantuuli ja aurinko", with normIt script. NormIt scales the input sound file's mean power to the desired value, this time being 60 dB, and returns a roughly rounded scaling factor as an output. It does not take into account the silent or very quiet parts of the sound, so it fulfils that requirement of accuracy. The rounding was disabled to get the scaling factors presentation in precision of several significant numbers. The scaling factors were calculated for all the synthesizers using their very own versions of the reference speech. NormIt was employed again, but this time it was forced to use the achieved scaling factors instead of the desired average power values. This way all the materials produced by each synthesizer could be quickly scaled to be equally loud on average. The masking noise was also normIt-processed to have mean power of 60 dB.

At this point, all the sentences were spoken into sound files, all having equal and comparable properties. Their sampling rate was 48 kHz, resolution of 16 bits, speaking rate roughly similar, and long-term power level 60 dB. The masker noise was of white type as described in section 4.1.3, stored in an own file, and sharing the same properties as the other sound files. The noise was band-limited to 0-8 kHz as described in section 4.1.3. Since the original sampling rates were various, in certain tts systems there is a possibility of some

sounds over 8 kHz being not covered by the noise. However, this should not be a serious problem, because the information carried by those high frequency sounds is minimal.

4.3 Test procedure

Performed test was based on HINT as described in section 3.3. The test contained totally 16 Finnish sentence lists spoken by four different TTS systems, and presented to each subject. Their task was to write down sentence they hear using a computer keyboard.

The studio in Helsinki University department of speech sciences was used as a listening test room. It has two separate rooms connected by a window: other for the test administrator with necessary equipment, and other for subject. Subject's room had been furnished with a table and a chair. On the table, there were a screen, a keyboard and a mouse for giving the answers. The sentences were played through a loudspeaker placed just behind the screen about 10 cm above it so the distance between the loudspeaker front panel and the listener was about 50 cm. Microphone-loudspeaker pairs were arranged so that the administrator and the subject could talk to each other.

The speech stimuli were reproduced by IBM ThinkPad T42 portable computer running Fedora Linux (kernel fc4 2.6.11) and Gnome desktop 2.10.0. Computer's soundcard was tested for its quality by playing white noise of band 20 Hz - 48 kHz. It was captured and briefly analyzed by another computer. The analysis showed a little attenuation at the highest frequencies, but in overall, the frequency response was straight by eye. ThinkPad's soundcard headphone output was attached directly to the subject's room active loudspeaker Genelec 1029A. This resulted to a little background noise always present, most likely due to grounding issues, but it was neglected as being minimal compared to intended noise levels. The ThinkPad screen, the keyboard, and the mouse were also connected to respective devices in the subject's room, so that the administrator and the subject had parallel controls over the system.

Overall listening volume was adjusted by playing the noise alone in its original level out of the computer. Its sound pressure level was measured from assumed head position of the listener, with Brüel&Kjær Presision sound-level meter 2203. It is a dated hand meter from the 60's, but still gives reliable measurements. The sound level was appraised to be suitable when adjusted to about 55 dB A-weighted. Noise louder than this combined with a loud sentence would have made the total volume unpleasant for the subject.

The listening test software applied in this work was called GuineaPig (Hynninen & Zacharov 1999), which is a generic software-based test platform for performing a wide range of subjective audio tests. Among the other test types, GuineaPig has ready procedures for SRT testing. After configuring the test, the administrator only needs to start a list

and the software handles the rest. It plays the first sentence of a list and the noise simultaneously to the subject and stops to wait for the written input. After the answer, the software evaluates if it is right. If not, the sentences' level is increased and it is presented again. This continues until the answer is correct. When the initial level of the list is reached with this procedure, the following items are presented. The levels are adjusted regarding to the answer by increasing it if answer is wrong and decreasing if it is correct. Subject's answers are stored in a result file, which additionally shows the correct answer for each sentence, SNR used, and a +/- tag about if the answer was correct.

GuineaPig settings in this work were as follows. The initial noise and speech levels were 0 dB and -10 dB, respectively. The speech level was adjusted by 4 dB steps for the first eight sentences in a list and with 2 dB steps for the last four. This was selected in hope of quick convergence at the beginning of a list, and precise convergence at the end of the list.

Fig.4.5 presents the subject's screen and a close-up picture of the answer window seen in the screen. After a sentence is presented, the answer window becomes active and the subject writes what he had heard and clicks the "Done" button.

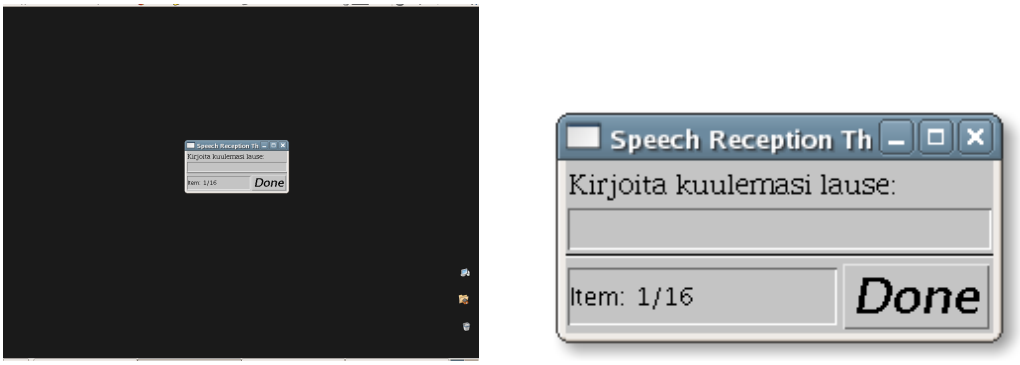

Figure 4.5: Subject's screen and a close-up picture of the answer window

All the subjects were native Finnish. None of them reported any hearing impairments. Everyone was volunteered and earned a movie ticket as a fee for the participation.

Each subject was familiarized to the test by going through one of the human-spoken lists (from Vainio et al. 1994) shortened to six items while the administrator was guiding alongside. The subjects were asked if they want to repeat the practise list, but none wanted. The actual test was performed so that the subject was left alone in the room, while administrator in his room started a list after another and turned on the subject screen (Fig. 4.5). Every now and then between the lists, the subject was asked how he is. A break was allowed in about a middle of the test. Each test took about one and half hour to complete, varying a little depending on subject's typing speed. This was considered as suitable time, as the subject reported it being quite long but workable. To stay in the schedule, a few subjects

were released when the time was full, although not all the lists were presented. However, the lists they managed to handle are still relevant results.

Typed answers were not case sensitive, so the capital letters at the beginning of the sentences and at proper names etc. could be written in low case. All the punctuation was also ignored.

4.4 MOS

MOS test was intended to follow the test suggested in, (Viswanathan & Viswanathan 2005). The first task in test preparation was to translate the questionnaire (see Appendix. A) into Finnish. The challenge is to get the questions, descriptions of questions, and the answer options designed so that the original concept remains, although not all the direct translations are available. Using common phrases used in Finnish MOS tests, the translations succeeded well, except a misinterpretation of the item "audio flow", which had turned to "lauseintonaatio (sentence intonation)". The resulting questionnaire is presented in Appendix A.

The text materials selected to the MOS test were short passages of speech of length about half a minute each. The passages concerned common announcements, e.g. news, weather reports, and public traffic timetables. Since there were 11 different items to judge in the questionnaire, the same amount of passages were prepared. The TTS systems were set to use the factor default speech properties settings, except a slight speed rate increase of tts3, to get it equal to the others. Each system read the passages into sound files, which were scaled with "normIt" script (described in the previous sections) to have the average volume of 60 dB. This helps to prevent the subjects to prefer some system due to its different volume level. In total, 44 speech files were prepared.

Six subjects attended the test, each being students in Helsinki University. They earned a course credit for the participation. Everyone was provided the questionnaires forehand to familiarize with. The same setting as in HINT was used: the listening room was equipped with the loudspeaker, chairs and tables and the administrator played the speech passages from the computer in the other room. All the subjects participated to the test simultaneously and they were randomly placed in the listening room.

Each subject had four questionnaires and a pen to mark their selection to those. Their task was to listen to a passage after another and always answer the next question. The passages were shuffled so that two consecutive passages from one TTS system should not occur, nevertheless all the items for each system were given a judgement at the end of the test. No further anchoring was used, the subjects had to appraise the items regarding to the samples heard during the test. There were approx. 10 s breaks between the passages for answering the question and reading the next one.

Chapter 5

Results and analysis

Totally 12 subjects participated to HINT. The aim was to present 16 lists for each, but because of a tight schedule, this was not possible completely. The results were achieved from 164 lists out of 192 intended. Results of the tests were stored as text files containing the following items to all the presented sentences: subject's answer, SNR of the sentence presentation level compared to the noise level and a sign about if the answer is correct. A script in Perl-language was prepared to parse the result files so that the desired SRT value for each list could be achieved. The script checked whether the last answer of a list (list item #16) was correct and added an extra SNR level to the parsed results as described in section 3.3. When the initial presentation level of a list was found by repeating the first item, the remaining sentences' presentation level converges towards the lists' SRT. This convergence should occur quickly, so only a few first items were neglected, and the items 5-17 were included into the calculation of SRT as an average:

$$SRT = \frac{SNR_5 + SNR_6 + \ldots + SNR_{17}}{13} \tag{5.1}$$

All the achieved SRT values are presented in the Table 5.1.

From the Table 5.1, the entire test results were calculated as an average for each TTS system. The individual result for list 1 was neglected for its being suspiciously good compared to the other results for the machine. The final results can be seen in Table 5.2, which contains the average SRTs and standard deviations from the average. For clarity, the results are also presented graphically in Fig. 5.1.

The next Section 5.1 briefly describes the exceptions occured during the test performance. Section 5.2 statistically confirms the applied task of combining all the results into one average value for each machine. Section 5.3 discusses the reasons for the results. Section 5.4 shows the results for the additional MOS test without a profound analysis of those.

Table 5.1: Results of the SRT test in dB. Rows indicate the list tested, columns are subjects. Lists are separated with respect to different machines.

List	Subj.1	2	3	4	5	6	7	8	9	10	11	12
1			-7.38									
2	5.85		4.62	5.23	0.92	4.62		3.38	-2.15	5.85	11.69	6.15
3	9.23	1.23		7.38	6.46		2.15	-0.62	9.54	1.85	4.00	0.31
4		1.54	6.46	2.77	4.92		3.38	2.77	4.31	3.38	6.15	3.38
5		3.69		5.54	10.46		12.31	9.85	0.00	10.77	4.92	6.46
6			-3.69	-4.31	-4.92		1.54	-3.69	-5.23	-2.15	-2.15	-6.77
7	-0.31		-3.69	-0.92	-1.54		3.69	-6.15	-3.69	-0.92	4.00	0.92
8	-0.62	-5.54	-5.23	-7.08	-4.62		-1.54	-5.23	-5.23	0.00	0.62	-7.38
9	2.15	0.92	-7.08	-4.31	-2.77	-2.15		-8.00	-8.62	-4.92	-1.23	-7.69
10												
11		-7.08	-6.77	-3.38	-4.62		-4.00	-5.23	-5.54	-4.00	4.00	1.23
12	-4.92	-7.08	-9.23	-8.62	-9.23		-8.62	-10.77	-9.23	-10.15	-7.38	-6.46
13	-7.69		-8.00	-9.54	-7.38	-8.31		-11.08	-4.92	-7.38	-6.15	-5.23
14	0.00		-3.08	-4.00	-3.38		-2.15	-8.62	-4.62	-5.23	-3.69	-0.92
15												
16				0.00								
17	-2.77	-5.54	-2.77		-2.15	-0.31		-1.23	-3.38	-4.00	-1.54	-0.62
18	-0.62	1.23	-3.08	1.85	-4.00		0.31	-3.38	-0.92	-0.92	0.31	3.38
19	-3.38	-3.38	-1.54	-4.62	-2.77		0.62	-4.62	-6.46	-3.08	1.23	2.77
20			0.00	-1.54	-4.31		1.23	-0.62	1.85	0.92	-0.92	-1.54

Table 5.2: Final results of HINT: the SRT values and the standard deviations

Tts system	1	2	3	4
Average SRT	4.89	-3.06	-5.82	-1.44
Std. dev	3.92	3.25	3.26	2.31

5.1 Observations during the test

Adjusting the initial level of a list by repeating and increasing the volume of the first sentence became complicated every now and then during the test. The subjects did not hear the sentence at low volumes and then, in the following trials, they might have got a false speech cue and understood a part of the sentence wrong. Most likely, the rest of the sentence was found to be complete nonsense regarding to the part that was wrongly heard. In such a case, the concentration easily focuses into hearing the "nonsense" part of the sentence in the following trials. Some subjects also reported that they might have heard something very strangely pronounced, but they just kept the answer what they wanted to think they heard. For these reasons, the initial volume of a list could grow unnaturally loud compared to the speech levels people are used to hear. In turn it could degrade the possibility to hear

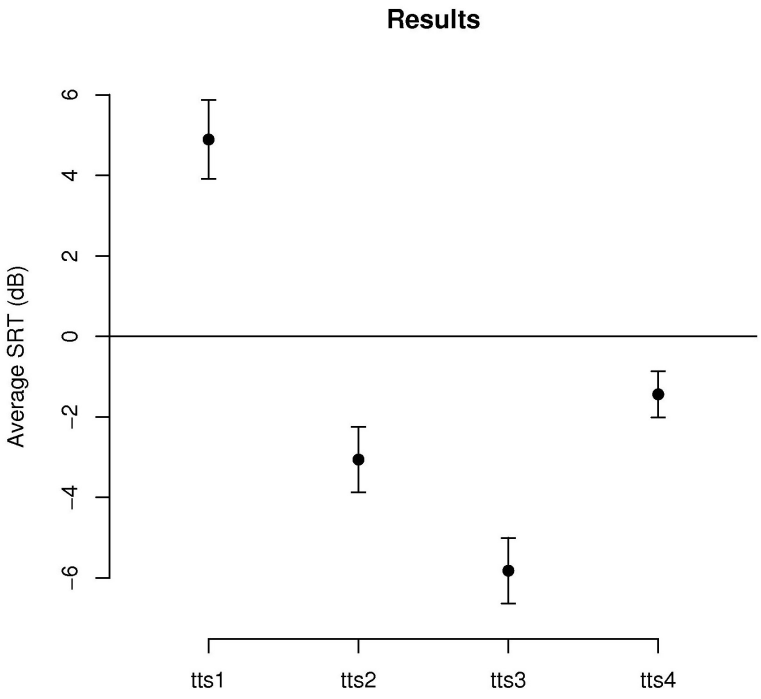

Figure 5.1: Final results of HINT as calculated in Tab. 5.2. The dot shows the average SRT bounded by the standard deviations.

the sentence correctly. Shouting in louder volume shapes also other voice properties than the level, so simple volume scaling like in this work might sound odd when done in large amounts. Very loud sounds are also unpleasant to hear. A few subjects mentioned that they needed to concentrate more on getting ready to tolerate the next too loud sentence, than to understand what it says. In addition, scaling the sentences too loud result nasty distortion due to exceeding the maximum representation level (Eq. 4.1).

To prevent the overdoing the initial adjustment, it was decided that the administrator could tell the correct answer to the lists' first sentence if the subject seemed not to figure it out before the volume went too high. This happened a few times when SNR was 14 dB and the sentence was still not heard. It was thought that this kind of administrative interfering would not affect the results. In these cases, the next few items in the lists were always correctly heard, which suggests that the subject had heard a false cue in the first sentence and was clinging on it.

Further, it was found that especially list 5 suffered from difficult first sentence: "Asiasta kertoi Espanjan radio". Machine's pronunciation could be poor or there could be something curious in the sentence, for example its several /s/ phones, that are totally covered

by wideband noise. Third sentence in that list contained a confusing proper name "Harry Hyökkääjä", which is a strange nickname and tends to be written as "Harri" by Finnish people. The latter form was accepted as a correct answer after a few subjects. All together, this sentence combination made three subjects fail the list, since it went too loud. The first sentence was traded to a different one from another list after subject no.9

A few other individual sentences were found commonly difficult. They contained proper names, which do not give any semantic cue about the sentence content. Instead, the proper name could almost be whatever, and therefore even one misheard phone can easily lead to a wrong answer. In addition, a few sentences were somewhat nonsense alone and would require proper context to be reasonable. These were however minority, and is assumed that the wrong answers should diminish in the total amount of answers.

5.2 Statistical analysis

The main question in the study is "Is there a difference in SRT values between different machines?" It can be examined in terms of the average SRT values for each machine, but there is a need for more careful inspection of the results to get a reliable answer.

Mainly two factors might affect individual machine's results. First, the lists might be unequal in difficulty although their human-made recordings are similar. This is crucial, since all the machines were reading the same lists associated to them during the test and if there is one more difficult list among others, the corresponding machine will suffer from it. Secondly, the subjects are different. It is assumed that everyone will have equal hearing capabilities and thus the same SRT, but in practise, this is not true. However, if one subject is biasing the results by achieving noticeable higher or lower values than others, most likely all the machines will be affected by this. It means that the results in general will be pushed to one direction, but none of the machines will suffer individually. The limited mount of lists per machine completed by one subject (maximum four) do not even give enough data to make reliable conclusions about individual subject performance.

To find out if there is difference between the lists, the analysis of variances (ANOVA) as described in (Milton & Arnold 1995) was planned. It is a method where the total variation of results is partitioned into components that are likely to represent the sources of variation. In this case, the components are lists. Assumptions when performing ANOVA are that each list's results are normally distributed and the variance in each list results is the same. Normality assumption was accepted as is. Because ANOVA uses F-test statistic, which is sensitive to the violation of the assumption of equal variances, that assumption still needed a confirmation.

Single spare list results (lists 1 and 16) were not included in these calculations, since one

result/list does not give much information about those lists themselves.

This section, or the procedure of statistical analysis of the results, is structured into the following subsections. 5.2.1 describes the test to confirm that the variance of the results for each list is similar enough. It is the prerequisite for performing ANOVA in 5.2.2, which confirms the similarity of the results of a list within a TTS system. That is, in the Table 5.1 the results of one row should not differ radically from the other rows within the section of a system. In 5.2.3, the statistical significance of the difference between each system's results is confirmed.

5.2.1 Testing equality of list result variance

As a preliminary work to ANOVA, the hypothesis

$$\begin{aligned} H_0 &: \sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2 \\ H_1 &: \sigma_i^2 \neq \sigma_j^2 \end{aligned}$$

was tested with *Bartlett's* test (see Milton & Arnold 1995), which tests the equality of all lists' result variances.

The sample variance S_i^2 of one list's results is an estimate of corresponding variance σ_i^2. It tells how well the results are clustering around list mean. It is calculated as

$$S_i^2 = \sum_{i=1}^{n} \frac{(X_i - \overline{X})^2}{n-1} \tag{5.2}$$

where X_i is an individual result, $\overline{X}$ is the mean of the list results and n is the sample size or number of results for the list in issue.

Under the assumption that H_0 is true, an unbiased estimate of total σ^2 is calculated as sample size -weighted mean of S_i's.

$$S_{tot}^2 = \sum_{i=1}^{k} \frac{(n_i - 1)S_i^2}{N-k} \tag{5.3}$$

where N is total sample size. Other variables needed for Bartlett's test are

$$Q = (N-k)\log_{10} S_{tot}^2 - \sum_{i=1}^{k}(n_i - 1)\log_{10} S_i^2 \tag{5.4}$$

$$h = 1 + \frac{1}{3(k-1)}\left(\sum_{i=1}^{k}\frac{1}{n_i - 1} - \frac{1}{N-k}\right) \tag{5.5}$$

$$B = 2.3026\frac{Q}{h} \tag{5.6}$$

If the variances between lists are very similar, Q will receive values close 0, but if they are different, Q will be big. Scaling Q with h and a constant gives a χ^2-distributed Bartlett statistic B, with $k-1$ degrees of freedom (γ). The variances are judged to be unequal if $B > \chi^2_{\alpha,k-1}$, where α is pre-selected significance level. Other way around, the probability can be studied as $1 - P[B]$ with the question: "What is the probability of variances being unequal when B and γ are given?" The higher B value is, the stronger evidence it gives to reject H_0. The area of accepting H_0 was selected to be 0-95%. Some selected χ^2 distribution values are presented in Table 5.4.

As an example, there are calculations for tts2 in Table 5.3. From these, the Bartlett statistic can be calculated using equations 5.4...5.6.

Table 5.3: Bartlett test calculations for tts2.

List	S_i^2	$\log_{10} S_i^2$	n_i
6	5.68	0.75	9
7	10.26	1.01	10
8	8.26	0.92	11
9	13.60	1.13	11

Achieved B statistics are 7.33, 1.67, 5.89 and 3.11, with respect to machine. Because $k = 4$, χ^2 distribution degree of freedom is 3. From Table 5.4 it can be seen that there is no need to reject H_0, except perhaps for tts1 and tts3, that have probabilities between 90-95% and 75-90%, that are quite high.

Table 5.4: χ^2 distribution $P[\chi^2_\gamma \leq t]$ for selected γ values

γ\F	0.050	0.100	0.250	0.500	0.750	0.900	0.950	0.975
3	0.352	0.584	1.21	2.37	4.11	6.25	7.81	9.35
11	4.57	5.58	7.58	10.3	13.7	17.3	19.7	21.9
14	6.57	7.79	10.2	13.3	17.1	21.1	23.7	26.1

Reasons for this behaviour can be found when inspecting the results more carefully. Tts3, in general, has results a few dB below zero. Only exception is one "+4" in list 11. Indeed, there are a few clear errors in writing in the answer files. By omitting this result, B value drops from 5.89 to 1.61.

Similarly for tts4, omitting lists 18 and 19 from subject 12, and list 19 from subject 11 makes B value drop from 3.11 to 0.49. There seems not to be any clear errors in writing in the answer files, which could explain these results.

Tts2 has overall average about -3 dB. Furthest items are list 9 from subjects 1 and 2, and list 7 from subject 11. Omitting these decreases B from 1.67 to 0.35. A common factor in

these results is subjects' fumbling in understanding the first sentences of the lists. It had led to relative high presentation levels and the situation was not compensated enough during the list. Similar reason could be also present above in tts4, subject 12, list 19.

Tts1 is problematic. It appears that the results' variance is huge. Several results are somewhere around 10 dB, which means that the presentation level has been very loud. It was discussed earlier that these levels are not very reliable any more. Especially list 5, that had first sentence obviously difficult for the machine, has received many of these high level results, but also a few that seem more reasonable. In addition, a few results are below zero, although overall mean is about 5 dB. Only list 4 has small variance, which suggests that the results are more systematic for that list. Omitting this list would make the B value drop from 7.11 to 0.098, so that the correlation between different lists seems to be good. Yet it is awkward to omit a list with systematic results in terms of getting more randomly distributed fit together.

Encouraging results in overall when comparing variances within each machine's lists led to further study of the issue. The listening test setting was the same for all the machines, so although each list could have a different mean, the variances should be similar. In other words, each lists results should cluster around its mean equally tight, no matter what the mean is. Bartlett's test was also performed combining all lists' variances. The procedure is the same, except now the χ^2 distribution degrees of freedom are different. For 16 lists, the degree of freedom is 15. B statistic gets value 25.86, which indicates poor correlation of variances, because the $P[B]$ value is between 95-97,5%. Removing tts1 from the calculations due to its poor performance even within its own lists, will decrease B to 15.09, but also degrees of freedom to 11. This is already promising, because the $P[B]$ gets an acceptable value 75-90%. In addition, above discussed suspicious results were removed: (list(L)7 subject(S)11), (L9 S1), (L9 S2), (L11 S11), (L19 S12). This decreased B to 7.08 and $P[B]$ to 10-25%.

Altogether, with a few individual exceptions, it could be trusted that each list's variance is the same. This was a good basis to start comparing list means.

5.2.2 Testing equality of list result mean

The mathematical model for this ANOVA experiment is

$$Y_{ij} = \mu_i + E_{ij} \tag{5.7}$$

$$Y_{ij} = \mu + (\mu_i - \mu) + (Y_{ij} - \mu_i) \tag{5.8}$$

$$Y_{ij} = \mu + \alpha_i + E_{ij} \tag{5.9}$$

where Y_{ij} is the result of jth subject and ith list, μ is the overall mean, α_i is deviation from overall mean due to effect of list i, and E_{ij} is random deviation from list i mean. If

one list is significantly different from the others, the factor that α parameter stands for, has pushed all the results of that list to same direction. Otherwise, all the lists' means μ_i are the same as the overall mean μ and $\alpha = \mu_i - \mu = 0$. Also, Eq. 5.7 says that each list's results deviates randomly around the overall mean.

Hypothesis to test is

$$\begin{aligned} H_0 &: \mu_1 = \mu_2 = \mu_3 = \mu_4 \\ H_1 &: \mu_i \neq \mu_j \end{aligned}$$

where H_0 stands for claiming that each lists' mean is the same, unless there is enough evidence that at least two means are not equal. Testing H_0 assumes that the random components E_{ij} are independent, normally distributed random variables, with mean 0 and variance σ^2.

Variables needed for ANOVA are

$$SS_{tot} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}..)^2 \tag{5.10}$$

$$SS_{list} = \sum_{i=1}^{k} n_i(\overline{Y}_i. - \overline{Y}..)^2 \tag{5.11}$$

$$SS_e = SS_{tot} - SS_{list} \tag{5.12}$$

$$MS_{list} = \frac{SS_{list}}{k-1} \tag{5.13}$$

$$MS_e = \frac{SS_e}{N-k} \tag{5.14}$$

$\overline{Y}_i.$ is the mean of ith list results and $\overline{Y}..$ is the mean of all results. SSs, the squared sums, are intermediate stages to calculate mean-squared estimates of list (MS_{list}) and random error (MS_e) effects to the results. In (Milton & Arnold 1995) it is shown that MS_e is actually an unbiased estimator of total variance σ^2 and $E[MS_{list}] = \sigma^2$ (also an unbiased estimator of σ) if α is zero, that is, there is no list dependent deviation from the mean.

From the variables above, an F-distributed statistic can be formed as

$$F_{k-1,N-k} = \frac{MS_{list}}{MS_e} \tag{5.15}$$

Its value will be one (1) if there is no list dependency or the variance is only due to random variation. Larger values indicate list dependency: it is more likely that at least one list has a different mean. Every machine has four (4) lists, so that the first degree of freedom is three(3). Second degree of freedom is dependent on total amount of results for each machine, being $41 - 4 = 37$ for three latter machines and $39 - 4 = 35$ for the first.

The confidence interval for ANOVAs was selected to be $P < 0.05\%$. With the confidence value and degrees of freedoms formed above, critical F values can be found by looking an F-distribution table. If F value achieved from an ANOVA exceeds its critical value, there is a reason to reject H_0. In this case, the critical F values are 2.87 for the first machine and 2.86 for the rest. ANOVA calculations for tts4 are presented in Table 5.5.

Table 5.5: ANOVA table for tts4.

Source of variation	Degrees of freedom	Sum of squares(SS)	Mean square(MS)	F
Lists	3	34.5	11.4	2.31
Error	37	182	4.92	
Total	40	216		

Running ANOVA for all machines gives the following F values with respect to machine: 1.79, 2.26, 11.2, and 2.31. Comparing these to the critical values shows that for tts1, tts2, and tts4, there is no reason to reject H_0. For tts2 and tts4, it is straightforward to say that the mean values of all lists are the same or that there is no list effect present. Thus, all the results of those machines can be combined as one set, from which further analysis can be performed. For tts1, there were problems in showing that the list variances are similar enough, which led to somewhat doubtful performance of ANOVA. However, ANOVA suggests that there is no list effect and therefore the results could be combined.

Tts3 highly exceeds the critical value. This suggests that there is at least one list with a different mean than others. Suspicious result (+4 dB) that was found when comparing variances was removed from further calculations. It however increased F to 12.5, which gives even stronger evidence to reject H_0. When taking into inspection the averages and variances of corresponding lists, it can be seen that lists 12 and 13 have significantly better performance over lists 11 and 14:

List	11	12	13	14
Average	-4.38	-8.34	-7.57	-3.57
Variance	5.98	2.94	3.50	5.73

There is no obvious way to overcome this. It seems that there are differences between lists with this machine. Easiest way to get the results comparable is just to neglect the study above and take a simple average. It seems that tts3 has all the list averages way below other systems, and therefore the results are combined for this system, despite the differences in list intelligibility.

With these arguments, the average SRT values and variances were calculated for each machine. The results in Table 5.1 are used as they are, except removing the list 1 from

subject 3. The final results can be seen in Table 5.2, which contains the average SRT and standard deviation taken as square root of the variance.

5.2.3 Testing the significance of difference in averages

From the results in the Table 5.2 it can be seen that tts1 separates clearly from the others. Yet, the standard deviation of the results for other systems overlap each other to some extent, so the question about if the results have any difference at all, remains. This was studied with Student's t-test for "the difference of two averages" as described in (Laininen 1998).

In the t-test, the hypothesis is that the averages of two datasets are equal. This is formulated as

$$\begin{aligned} H_0 &: \mu_1 - \mu_2 = 0 \\ H_1 &: \mu_i - \mu 2 \neq 0 \end{aligned}$$

The hypothesis is tested with a T-distributed t_0-statistic, which is calculated as

$$t_0 = \frac{\overline{x}_1 - \overline{x}_2}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, \nu = n_1 + n_2 - 2 \tag{5.16}$$

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} \tag{5.17}$$

Statistics t_0 were calculated separately for the difference between tts2 and tts3 as well as between tts2 and tts4. The distance of results between tts3 and tts4 is already so big, that they were considered as different. Found t_0 values were 3.26 for $\mu_{tts2} - \mu_{tts3}$ and 2.60 for $\mu_{tts2} - \mu_{tts4}$. The corresponding p-values for achieved t_0's were revised with statistical software. For $\mu_{tts2} - \mu_{tts3}$, the two-tailed p-value with $\nu = 80$ degrees of freedom, is 0.0016. It denotes a very high significance for to reject H_0. For $\mu_{tts2} - \mu_{tts4}$, the two-tailed p-value with $\nu = 81$, is 0.011. With ordinary significance level, 0.05, H_0 has to be rejected also in this case.

Altogether, it can be confidently claimed that the results are separable. Each system has a statistically distinctive SRT value.

5.3 Discussion

From HINT results, the TTS systems can be ranked into an order regarding to their intelligibility in noise. The average SRT values for each system are available in the Table 5.2 and in the Fig. 5.1. Most intelligible system is tts3, which produces speech that can be understood from almost 6 dB louder white noise. Other extremity is tts1, where the speech

needs to be almost 5 dB louder than the noise to be understandable. The systems tts2 and tts4 fall in between, tts2 being slightly more intelligible.

It should be noticed that despite the statistically distinctive results, in (Vainio et al. 1994) the sensitivity of the test is found to be around 1.5 dB in Finnish. That is, the true difference in intelligibility can be claimed only when the results differ more than the sensitivity is. When obeying this rule, the difference between tts2 and tts4 decreases minimal.

Reason for good test performance of tts3 is probably its good spectral fit into human hearing. Also of importance is the low-level synthesis method, which produces the speech from parameters, instead of other systems, that use concatenation of pre-recorded speech samples. As summarized in (Tokuda et al. 2002), the HTS system has the advantages of speech being smooth and stable, while having a tone of vocoded speech. This drawback affects the speech naturalness, but as suggested in (Viswanathan & Viswanathan 2005), naturalness and intelligibility can be assumed as two independent factors of speech quality. Therefore, while HTS system might sound unnatural, it indeed is superior in intelligibility over the competitors, which perhaps are more natural.

Well-modelled supra-segmental properties of speech in parametric synthesis support the assumption of proper prosody model affecting positively to the intelligibility or comprehension. As seen in Chap. 4.1, tts3 has the finest intonation curve following closely to the intensity curve, effectively emphasizing segments of speech. It results to naturally stressed, or even exaggerated speech, which gives the listener good prosodic cues about the message, although some segments are missed.

Tts2 and tts4 results are likely to represent those of concatenative systems in general. It is believed that tts4 performs worse because of the concatenation distortion, and possible other signal processing distortions, which occasionally are even audible. The distortions are also seen in the long-time spectrum of the speech. Concatenation of segments, which do not fit so well together, produces unwanted frequency components that spread over the spectrum. When scaling the speech volume to the same level as others have, the distortion components will also be emphasized and the informative parts are left more silent. Other reason for the results is probably the flat intensity pattern of tts4, although the intonation seems more natural than in tts3. The variance of results of tts4 is relatively small compared to the others throughout the test. This suggests that the speech tolerates noise relatively well to some extent, but after that, the intelligibility suddenly drops. It is understandable because the spectrum and the intensity are flat, so the speech will also be covered in the flat noise when it is loud enough.

Tts1 performed too badly in the test. As described earlier in this chapter, the results are almost beyond the practical limits. Masking noise should have been more silent if appropriate results were needed. In the initial sentences of HINT lists, the listeners often provided

a wrong word and kept that while the sentence was repeated with increasing volume. The obvious defect in the system is bad prosody, which does not support the listener to understand the message. It is possible that the stress of the words was miss-aligned. For example, if in natural speech, the stress is placed on the first syllable of the words, and the synthesis places the stress on second syllable, the word is found to begin from the wrong place in the noise-interfered sentence.

Results suffer from rather large variance and thus large standard deviation, its being $\sigma > 3dB$ for systems tts1, tts2 and tts3, while (Vainio et al. 1994) achieved standard deviation of $\sigma = 1.61dB$ for Finnish lists. The reason for this most likely is the differences in the sentences, others being more difficult to understand than others are. In the original HINT, this problem was overcome by studying the sentence difficulty forehand, and scaling the respective sentences in volume, which was omitted in this work.

In overall, speech synthesis seems to be intelligible in noise that is even a few decibels louder in average volume. The human hearing is so dedicated to the signal patterns of speech, that it can be rebuilt from pieces and contextual cues, although not everything is heard. The study shows that synthesis is successful in producing such patterns, even so that the best synthesis resulted to SRT of almost -6 dB, while (Vainio et al. 1994) achieved SRT about -2.5 dB in natural Finnish voice. However, further comparison of the results is not preferable, especially due to different types of background noises used.

5.4 MOS results

Results achieved from additional MOS test are only indicative. The test setting was deficient because of small sample size (six persons) and lack of proper anchoring. Unfortunate mistranslation of item "audio flow" into "sentence intonation" leads to neglecting the whole item, because the new version has not been carefully examined to represent an appropriate feature in speech. In addition, the concept of intonation might have been unclear to the subjects, so there is no assurance that the answers concern the same thing.

The results of the MOS test can be seen in the Table 5.6. The answers were enumerated so that the option describing the best quality or describing least degradation received highest value. The results are calculated as averages of each subjects' results. The items are arranged into the factors of "naturalness" and "intelligibility", as suggested in (Viswanathan & Viswanathan 2005). The scales for each item are from 1 to 5, except for the item on acceptance, which has the scale from 1 to 2.

In Table 5.7, the averages of each factor are calculated. The results are also calculated for the total MOS value, including all the items, except the item on acceptance.

The interesting observation in the MOS results is the good performance of tts4, while

Table 5.6: Results of the MOS test arranged into factors of naturalness and intelligibility. The scales are from 1 to 5, except for the item on acceptance, where the scale is from 1 to 2.

	tts1	**tts2**	**tts3**	**tts4**
Naturalness				
Naturalness	2.29	1.57	3.71	3.43
Ease of listening	1.86	2.57	2.86	4.14
Pleasentness	1.86	2.43	2.86	2.57
Intelligibility				
Listening effort	3.29	3.86	4.00	4.14
Pronunciation	2.14	2.29	3.43	4.14
Comprehension	3.86	4.57	4.43	4.29
Articulation	3.00	3.71	4.14	4.43
Speaking rate	4.00	4.71	4.29	4.00
Overall quality				
Overall impression	2.57	2.43	3.14	3.71
Acceptance (1-2)	1.00	1.00	1.71	1.71

Table 5.7: MOS test results combined to the factors and summarized to total MOS value, excluding the item on acceptance.

	tts1	**tts2**	**tts3**	**tts4**
Naturalness	2.00	2.19	3.14	3.38
Intelligibility	3.26	3.83	4.06	4.20
Total MOS	2.76	3.13	3.65	3.87

other systems' results follow the order of HINT. This suggests that the unit-selection type synthesis is highly natural and even very intelligible when heard in good conditions, but will fall apart when there is enough interfering noise present.

The biggest difference between tts3 and tts4 is in item on ease of listening. This could be due to HTS-type synthesis' vocoded tone, which is seen irritating in long-time listening. In addition, the worse performance in intelligibility-related items in MOS compared to HINT places a question about if the parametric synthesizer overdoes the speech: the shaped formants are well perceptible in noise, but become unnatural and thus not preferred in absence of noise.

Chapter 6

Conclusions and future work

In this work, HINT that originally was developed to measure the degree of hearing impairment, was prepared and performed to evaluate the intelligibility of current Finnish TTS systems. It was found that the test effectively could distinguish the extremities of systems by the means of how well they can tolerate the noise interference and still be understandable.

The system tts3 showed its superiority over the competitors in the intelligibility. Tts2 and tts4 took the following positions, from which the test resolution can just barely tell the predominance of tts2. Tts1 was left far away from the others, being almost too difficult to understand for the test setting used. The success of tts3 predicts a hopeful future for the parametric methods in speech synthesis, especially that of HMM models. However, the size of the sample, namely four systems, does not provide a reliable reason to claim that the concatenative systems were worse in overall. Therefore, it would be interesting to compare these results with corresponding results in other languages and with other TTS systems. In addition, it would be interesting to see if the concatenative systems would perform better if they were filtered to emphasize similar frequencies that tts3 is found to do.

HINT test in the form described in this work is hardly suitable for general testing of TTS systems, mostly due to large resource requirements. The test performance for a subject takes about one and half hour to complete, and if everything goes well, 16 SRT values can be achieved, which is not very much compared to the time used. Efficiency of time usage could be enhanced by developing the test procedure, perhaps by abandoning the computer aided answering. The easiness of administration of the test is ostensible, since the administrator needs to be observant of the test progress and change the sentence lists to play. The subjects' task is monotonous: the typing is not a natural way to communicate and is thus slow and tedious. Hand written answers would not solve this problem, as they also are rather slow to manage. A potential test procedure could be that the subject tells

the administrator spoken answers, from which the administrator could conclude the level of the next sentence, and play it. This would make the test quicker to perform, but on the other hand, it would introduce other problems, such as effect of administrator's false heard correct answers.

The sentence lists of HINT may need further modification still, because relatively large variances of the results suggest that not all the sentences are equally difficult. Whatever the reason is, the scaling of the sentences before the test seems to be an appropriate operation, unless the sentences could be divided into equal predictability groups in some manner. There could be a need for proper analysis on distinction of semantically predictable and unpredictable sentences. The latter category would definitely contain the sentences with proper names, and the sentences that would need correct context to be reasonable. In addition, some proper names found in the lists are currently familiar to some people, but can be unfamiliar to the others, and be forgotten in near future. For example, the names of sportsmen and politicians are such. Including sentences of this kind to the test should be considered with care.

The list effect could be further reduced by mixing up the lists so that different subjects will get the same lists spoken by different systems. If there were lists that are more difficult or easy, the effect of those would be shared with all the system, instead of just one.

In all SRT testing, the volume scaling of the speech files should be done so that the maximum dynamics is maintained. When initially scaling to e.g. 60 dB, as was done in this work, the volume increase during the test also makes the quantization noise louder. A better way would be to scale all the files to the maximum possible volume that is common to all the speakers. During the test, the attenuation of the files should be altered instead of gain, since it only can diminish the quantization noise. With attenuation control of the maximum dynamics files, exactly the same test setting can be arranged with smaller amounts of unwanted noise.

A proper selection of the masking noise is essential. This work incorporated white noise in frequency band 0-8 kHz, from which some high frequency components of speech can leak unmasked. Most likely, their minor existence does not affect the results, but there is no confidence about that. A secure method would be to limit both the speech and the noise to the same band, but in such manner, that the sound of the systems would not suffer. The long-term spectrum shaped noise could also be worth of considering.

The question about the possible synthetic "super speech" that is more intelligible than natural, remains. Now there are encouraging results supporting the high intelligibility of synthetic speech in noise, but there should also be the results for natural speech with the same test setting. However, these ideas raise the question about the definition of the term "intelligibility". Is the speech that has better SRT value in noise always more intelligible

that another speech? Could the other speech be more intelligible with other kind of test setting? An appropriately performed MOS test could shed light on these questions.

In every case, the intelligibility is an important part of the perceived speech quality, which is a difficult concept to measure. In the very end, that is because of the lack of knowledge about what happens inside the heads of people, when they are communicating with speech.

Bibliography

Donovan, R. E. (1996), Trainable speech synthesis, PhD thesis, Cambridge University.

Donovan, R. E., Franz, M., Sorensen, J. S. & Roukos, S. (1999), Phrase splicing and variable substitution using the ibm trainable speech synthesis system, *in* 'Proc. of ICASSP'.

Donovan, R. & Eide, E. (1998), The ibm trainable speech synthesis system, *in* 'Proc. of ICSLP'.

Dubno, J., Dirks, D. & Morgan, D. (1984), 'Effects of age and mild hearing loss on speech recognition in noise', *J. of the Acoustical Society of America* **76**(1).

Egan, J. (1948), 'Articulation testing methods', *Laryngoscope* **58**, 955–991.

Fairbanks, G. (1958), 'Test of phonemic differentation: The rhyme test', *J. of the Acoustical Society of America* **30**(7).

Fant, C. (1970), *Acoustic theory of speech production*, Mouton, Netherlands.

Festival home page (2006), Internet. http://www.cstr.ed.ac.uk/projects/festival/.

Gelfand, S., Ross, L. & Miller, S. (1988), 'Sentence reception in noise from one versus two sources: Effects of aging and hearing loss', *J. of the Acoustical Society of America* **83**(1).

Goldstein, M. (1995), 'Classification of methods used for assessment of text-to-speech systems according to the demands placed on the listener', *Speech communication* **16**, 225–244.

House, A. S., Williams, C., Hecker, M. H. L. & Kryter, K. D. (1963), 'Psychoacoustic speech tests: A modified rhyme test', *J. of the Acoustical Society of America* **35**(11).

Hunt, A. J. & Black, A. W. (1996), Unit selection in a concatenative speech synthesis system using a large speech database, *in* 'Proc. of ICASSP'.

Hynninen, J. & Zacharov, N. (1999), Guineapig - a generic subjective test system for multichannel audio, *in* 'Proceedings of the Audio Engineering Society; 106th International Convention'.

Imai, S. (1983), Cepstral analysis synthesis on the mel frequency scale, *in* 'Proc. of ICASSP'.

ITU (1994), 'A method for subjective performance assessment of the quality of speech voice output devices', International Telecommunication Union. ITU-T recommendation, p.85.

Jekosch, U. (1992), The cluster-identification test, *in* 'Proceedings of ICSLP 92'.

Kalikow, D. N., Stevens, K. N. & Elliot, L. (1977), 'Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability', *J. of the Acoustical Society of America* **61**(5).

Karjalainen, M. (1999), *Kommunikaatioakustiikka*, Otamedia oy.

Klatt, D. (1987), 'Rewiev of text-to-speech conversion for english', *J. of the Acoustical Society of America* **82**(3).

Kraft, V. & Portele, T. (1995), 'Quality evaluation of five german speech synthesis systems', *acta acustica* **3**, 351–365.

Laininen, P. (1998), *Todennäköisyys ja sen tilastollinen soveltaminen*, Yliopistokustannus/Otatieto.

Lemmetty, S. (1999), Review of speech synthesis technology, Master's thesis, Helsinki University of Technology.

Logan, J. S., Greene, B. G. & Pisoni, D. B. (1989), 'Segmental intelligibility of synthetic speech produced by rule', *J. of the Acoustical Society of America* **86**(2).

Masuko, T. (2002), HMM-based Speech Synthesis and Its Applications, PhD thesis, Tokyo institute of technology.

Milton, J. S. & Arnold, J. C. (1995), *Introduction to probability and statistics*, 3 edn, McGraw Hill.

Nilsson, M., Soli, S. D. & Sullivan, J. A. (1994), 'Development of the hearing in noise test for the measurement of speech reception thresholds in quiet and in noise', *J. of the Acoustical Society of America* **95**(2).

Palo, P. (2006), Review of articulatory speech synthesis, Master's thesis, Helsinki University of Technology.

Plomp, R. & Mimpen, A. M. (1979), 'Improving the reliability of testing the speech reception threshold for sentences', *Audiology* **18**, 43–52.

Polkosky, M. D. & Lewis, J. R. (2003), 'Expanding the mos: Development and psychometric evaluation of the mos-r and mos-x', *International Journal of Speech Technology* **6**, 161–182.

Rohde, D. (2006), 'Splitit program home page', Internet. http://tedlab.mit.edu:16080/ dr/SplitIt/.

Smith, J. O. (2002), 'Digital audio resampling home page'. http://www-ccrma.stanford.edu/~jos/resample/.

Snack home page (2006), Internet. http://www.speech.kth.se/snack/.

Tokuda, K., Kobayashi, T. & Imai, S. (1995), Speech parameter generation from hmm using dynamic features, *in* 'Proc. of ICASSP'.

Tokuda, K., Zen, H. & Black, A. (2002), An hmm-based speech synthesis system applied to english, *in* 'IEEE Speech Synthesis Workshop'.

Vainio, M., Suni, A., Järveläinen, H., Järvikivi, J. & Mattila, V.-V. (1994), 'Developing a speech intelligibility test based on measuring speech reception thresholds in noise for english and finnish', *J. of the Acoustical Society of America* **95**(2).

Viswanathan, M. & Viswanathan, M. (2005), 'Measuring speech quality for text-to-speech systems: development and assessment of a modified mean opinion score (mos) scale', *Computer Speech and Language* **19**, 55–83.

Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T. & Kitamura, T. (1999), Simultaneous modeling of spectrum, pitch and duration in hmm-based speech synthesis, *in* 'Proc. of Eurospeech'.

Appendix A

MOS questionnaires

Overall impression

How do you rate the quality of the sound of what you just heard?

- O Excellent
- O Good
- O Fair
- O Poor
- O Bad

Listening effort

How would you describe the effort you were required to make in order to understand the message?

- O Complete relaxation possible; no effort required
- O Attention necessary; no appreciable effort required
- O Moderate effort required
- O Considerable effort required
- O No meaning understood with any feasible effort

Pronunciation

Did you notice any anomalies in pronunciation?

- O No
- O Yes, but not annoying
- O Yes, slightly annoying
- O Yes, annoying
- O Yes, very annoying

Speaking rate

The average speed of delivery was:

Just right **O**

Slightly slow **O** Slightly fast

Fairly slow **O** Fairly Fast

Very slow **O** Very fast

Extremely slow **O** Extremely fast

Voice pleasantness

How would you describe the voice?

- O Very pleasant
- O Pleasant
- O Neutral
- O Unpleasant
- O Very unpleasant

Voice naturalness

Did the voice sound natural?

- O Very natural
- O Natural
- O Neutral
- O Unnatural
- O Very unnatural

Ease of listening

Would it be easy to listen to this voice for long periods of time?

- O Very easy
- O Easy
- O Neutral
- O Difficult
- O Very difficult

Comprehension problems

Did you find certain words hard to understand?

- O Never
- O Rarely
- O Occasionally
- O Often
- O All of the time

Articulation

Were the sounds distinguishable?

- O Yes, very clear
- O Yes, clear enough
- O Fairly clear
- O No, not very clear
- O No, not at all

Acceptance

Do you think that this voice could be used for an interactive telephone system or a handheld device?

- O Yes
- O No

Figure A.1: MOS questionnaire as suggeested by Viswanathan & Viswanathan (2005)

SUOMENKIELISTEN SYNTEESIJÄRJESTELMIEN EVALUOINTI

8.12.2005 Mies / Nainen Ikä: ___

Yleisvaikutelma
Kuinka arvioisit kuulemasi näytteen laatua?
___ Erinomainen
___ Hyvä
___ Kelvollinen
___ Puutteellinen
___ Huono

Kuuntelemisen vaivattomuus
Jouduitko ponnistelemaan ymmärtääksesi viestin?
___ Pystyin rentoutumaan täydellisesti, eikä kuunteleminen vaatinut ponnisteluja.
___ Jouduin kiinnittämään huomioni, mutta en kuitenkaan erityisesti ponnistelemaan.
___ Kuunteleminen vaati vähäisiä ponnisteluja.
___ Kuunteleminen vaati huomattavia ponnisteluja.
___ En ymmärtänyt viestiä ponnisteluista huolimatta.

Ääntäminen
Huomasitko ääntämisessä poikkeavuuksia?
___ En.
___ Kyllä, mutta ne eivät olleet ärsyttäviä
___ Kyllä. Koin ne hieman ärsyttävinä.
___ Kyllä. Ne olivat ärsyttäviä.
___ Kyllä. Ne olivat todella ärsyttäviä.

Puhenopeus
Puhenopeus oli keskimäärin:
___ Juuri oikea
___ Hieman nopea tai hieman hidas
___ Melko nopea tai melko hidas
___ Hyvin nopea tai hyvin hidas
___ Erittäin nopea tai erittäin hidas

Miellyttävyys
Kuinka kuvailisit äänen miellyttävyyttä?
___ Hyvin miellyttävä
___ Miellyttävä
___ Siltä väliltä
___ Epämiellyttävä
___ Hyvin epämiellyttävä

Luonnollisuus
Kuinka arvioisit äänen luonnollisuutta?
___ Hyvin luonnollinen
___ Luonnollinen
___ Siltä väliltä
___ Epäluonnollinen
___ Hyvin epäluonnollinen

Lauseintonaatio
Kuinka arvioisit äänitteen intonaatiota?
___ Hyvin sujuva
___ Sujuva
___ Siltä väliltä
___ Katkonainen
___ Hyvin katkonainen

Pitkäkestoinen kuuntelu
Olisiko helppoa tai vaikeaa kuunnella ääntä pitkäkestoisesti?
___ Hyvin helppoa
___ Helppoa
___ Siltä väliltä
___ Vaikeaa
___ Hyvin vaikeaa

Kuullun ymmärtäminen
Olivatko jotkut sanat mielestäsi vaikeita ymmärtää?
___ Ei lainkaan
___ Harvoin
___ Satunnaisesti
___ Usein
___ Jatkuvasti

Artikulaatio
Olivatko äänteet erotettavissa?
___ Hyvin selvästi
___ Selvästi
___ Siltä väliltä
___ Epäselvästi
___ Hyvin epäselvästi

Hyväksyttävyys
Voidaanko ääntä mielestäsi käyttää esim. tiedotusjärjestelmissä?
___ Kyllä
___ Ei

Figure A.2: MOS questionnaire translated to Finnish

Appendix B

SRT lists

List 1

Ohi ei silläkään ollut asiaa.
Tyttö menehtyi välittömästi.
Jarmo sopii porukkaan mainiosti.
Argentiina on ottelun suosikki.
Entä jos Lipponen saa pojan?
Ratkaisevaa on äänestäjien määrä.
Turvallisuudesta kannattaa maksaa.
Sopiiko tällainen kansojen kotiin?
Talo valmistuu toukokuun lopussa.
Ateenan reissu ei ollutkaan turha.
Silloin yhdistystä veti Aho.
Tilanteet vaihtelevat vuosien myötä.
Sitä kukaan Unkarissa ei halua.
Oluiden myynti laski hieman.
Opettajat elävät sen keskellä.
Tarkkoja suunnitelmia ei ole.

List 2

Haaste on tässäkin mielessä uusi.
Menestystä ei juuri ole tullut.
Aivan oma lukunsa ovat pilvet.
Taiteilija itse jäi kotiin.
Hakemuksia on aina kymmeniä.
Opettaja jatkaa sitten työtäni.
Bergin rooli voi olla tänään iso.
Turkuun joukkue matkustaa perjantaina.
Tuon kamppailun Suomi hävisi.
Toisen karsinnan voitto meni Ranskaan.
Rinteessä vaiva ei ole tuntunut.
Onnea ei löydykään kaupungeista.
Anna katseli kaaosta ympärillään.
Virtanen huusi Kakkoselle.
Yhtiön tulos parani selvästi.
Presidentti myös johti puhetta.

List 3

Apua, sängyssäni on sika.
Onneksi Naganoon on vielä aikaa.
Kyse on nimenomaan sankarista.
Ohjelma uusitaan keskiviikkona.
Ohjeet tulevat Pietarista.
Norjalla on jo käytössään Paseja.
Pelien alussa se vähän tuntuu.
Jäljelle jäi vain hyviä muistoja.
Hitlerin puvun alta löytyy nainen.
Silloin demokratiakin saattaa toimia.
Demarien rivit ovat selvät.
Ramsaun kisasta on luvassa tiukka.
Puhetta johtaa Ilkka Kuusisto.
Työ kantaa kummallista hedelmää.
Minkälaista luonnetta se vaatii?
Vaiva kuitenkin uusii herkästi.

List 4

Osaltaan myös kaupunki on mukana.
Kotimme oli turvallinen paikka.
Asetta ei onneksi löytynyt.
Tavallaan lomasta käy myös teatteri.
Mies puhuu, naiset kuuntelevat.
Ongelmia ei silti ole tullut.
Itse tauti ei kuitenkaan parane.
Rata oli suhteellisen helppo.
Tiistaina vesi jäi metrin päähän.
Osa vuoroista jäi kokonaan väliin.
Esityksiä on joulukuulle asti.
Ehtoja voi saada jatkossakin.
Kokoomuksessa tämä tiedetään.
Hirvi muistuttaa niin paljon hevosta.
Anottiinko vain Suomesta turvaa.
Levykö ajoi bändin tien päälle?

List 5

Asiasta kertoi Espanjan radio.
Osaamista kyllä löytyy Suomesta.
Harry Hyökkääjä hoitaa verotuksen.
Muutos astui voimaan vuoden alusta.
Vaikutteet tekevät kaltaisekseen.
Turnaus lähti sen jälkeen hyvin käyntiin.
Tarkastellaan perheen malleja.
Apua, keittiön kaappi putosi.
Torstaina on vuorossa Berliini.
Ottelut jatkuvat sunnuntaina.
Tennis on koko elämäni.
Näin ei todellisuudessa ole.
Markuksen makkara on jo valmis.
Samaa mieltä on Johansson itsekin.
Kuitenkin isäni rikkoi lupauksen.
Intia oli yhä enemmän yksin.

List 6

Eläkkeestä maksetaan veroa.
Eevalla on ongelmia miesten kanssa.
Naiset yhdistävät heimoja.
Afrikka on opettanut muutakin.
Juttu oli ohi sekunnissa.
Samalla muuttui siviilien asema.
Kerrankin saa tehdä käsillään itse.
Tehtaan tuotantoon palo ei vaikuta.
Vapon omistaa Suomen valtio.
Kysymys oli vain sekunneista.
Maksu sisältää vuokran ruuan ja hoidon.
Lomautuksiinkaan ei ole tarvetta.
Töitä riittää, kommentoi Jortikka.
Kyselyjä tulee laidasta laitaan.
Pätevinhän sinne piti valita.
Aina joku odottaa kirjettäsi.

List 7

Enemmistö on silti vielä miehiä.
Rytmi sammuu ja liike lakkaa.
Mikä keskustan ryhmässä muuttuu?
Tampereen kokous jatkuu lauantaina.
Venäjänkin luottamus on kunnossa.
Kofi Annan avasi kisat.
Puhettakin on jo riittänyt.
Tuolloin lopulliset luvut selviävät.
Oopperan tuottaa Risto Hirvonen.
Orkesteria johtaa Juha Kangas.
Yhtä selvä kakkonen on Suomi.
Tontit myydään tarjousten perusteella.
Eilinen päivä keskittyi Tallinnaan.
Asut ovat aina viimeisen päälle.
Vaara lähti maailmalle Porista.
Esitys on edelleen voimassa.

List 8

Ero tuli voimaan heti torstaina.
Lasten parlamentteja on muuallakin.
Enimmäkseen tapahtuu jälkimmäistä.
Sitten on edessä uudet haasteet.
Luku oli koko ajan kasvussa.
Kilpailut päättyvät tiistaina.
Jäljet nimittäin pelottavat.
Ainakaan raiskaus ei ole lievää.
Stefan keksi siirtää pihan katolle.
Vastaava malli voi syntyä myös Turkuun.
Onneksi on muitakin kisoja.
Minne joutuvat vanhat tietokoneet?
Näin ei todellakaan tapahtunut.
Ikkunan alta lähtevät junat.
Oikeus hylkäsi vetoomuksen.
Prosessi oli sen verran rankka.

List 9

Erä oli jo tulossa kotiin.
Prosessissa syntyy alkoholia.
Tuotannosta menee kolmannes Ruotsiin.
Ahvenanmaa on siis veden maa.
Sopimus menee vielä valtuustoon.
Venäjä kuitenkin kiisti väitteet.
Edellinen on Helsingissä.
Huomenna yritetään uudestaan.
Hirvi voi nimittäin ampua takaisin.
Ajaako kauppias pelaajan etua?
Lomautuksia palosta ei aiheudu.
Järjestö valitti päätöksestä.
Loimaan heikkous on kokemuksen puute.
Nykyiset tilat kaipaavat käyttäjää.
Niinistön vallallakin on rajansa.
Markkinointi oli hyvin vähäistä.

List 10

Tampere on vuorossa perjantaina.
Alkoholin käyttö on toinen juttu.
Irtosihan se hymy viimein.
Hakemuksia saattaa vielä tulla.
Ainekset ovat Sudessa oivat.
Onnistuminen riippuu itsestämme.
Myöhemmin näistäkin toinen kuoli.
Antti Laakso maalasi kahdesti.
Kirjassa mainitaan myös harrastukset.
Ottelu oli kuitenkin tiukka.
Sauli Niinistöstä pätee sama.
Latvia on kallis maa turistille.
Näin Agenda jää Emun varjoon.
Serbit jatkavat pakoaan kaupungista.
Onnettomuuden syytä selvitetään.
Joukkueen johtaja on Jyrki Uotila.

List 11

Silti on syytä olla huolissaan.
Lasi vei miehen kuitenkin mukanaan.
Suunnitelma tehdään paperille.
Pankkien tulos parani edelleen.
Siellä nousi paksua savua ilmaan.
Koneet kiinnostavat viljelijöitä.
Seppälä ei pidä dopingista.
Luvut kertovat hankkeen laajuudesta.
Kuka hautaa lemmikin ja minne?
Harjoituksen sisältö on muuttunut.
Jakso ei ole yhtenäinen.
Arvot erottavat myös naisia.
Asia voi johtaa jopa äänestykseen.
Nestettä pitää nauttia runsaasti.
Kritiikki on selvästi harmittanut.
Kalevi käy myös sairaalassa.

List 12

Finaalissa Suomi kohtaa Venäjän.
Kumpikin istui hyvin rooliinsa.
Tehoa ei Tikkanen ole löytänyt.
Brysselistä on tullut uusi Kreml.
Poliisi etsii tekijöitä.
Seuraukset voivat olla kohtalokkaat.
Kotiin pitää mennä kaupan kautta.
Sävellys todella kosketti.
Menneisyydestä voidaan oppia.
Ajatus ei ole mitenkään uusi.
Sveitsiin joukkue matkustaa perjantaina.
Ongelmana on hoidon kallis hinta.
Kunnan rajako ihmisen määrittää?
Aina on joku tulossa tilalle.
Presidentti myös vahvistaa lait.
Palo sai alkunsa tämän jälkeen.

List 13

Verkkoa rakennetaan vähitellen.
Ajatus pysyi hyvin kasassa.
Aurinko lämmittää mukavasti.
Seuraajia on jo ilmaantunut.
Selkeä sydän sijaitsee kaupungissa.
Vahinkohan on jo tapahtunut.
Lajusen taktiikka oli selvä.
Ottelu oli todella vauhdikas.
Pelätään että sairautemme tarttuu.
Yksi hävittää, toinen nappaa talteen.
Ehkä Ruotsi taipui juuri silloin.
Maaliskuu on Rantasen mukaan kylmä.
Pieni näyttämö on haasteiden paikka.
Vieraskin tuntee tulevansa kotiin.
Tampere sai jo toisen myymälän.
Kuoro laulaa yleisön edessä.

List 14

Islannissa se pitää paikkansa.
Simo taistelee nyt terveydestään.
Ammattiin halutaan hyvä aines.
Tallinna loisti tässäkin kärjessä.
Merja Kääriäinen kävi tekstit läpi.
Edut menevät joskus ristiin.
Silloin sen iho menee rikki.
Odotuksetkin ovat kovat.
Omaisuuttakaan ei juuri ole.
Musiikki on rakas harrastuksemme.
Formulasta loppui vain veto kesken.
Hattu on myös aivan oma maailmansa.
Jäljet johtavat suoraan tornin alta.
Opettajista puuttui vain yksi.
Niin painokin pysyy hallinnassa.
Hahmo ei kuitenkaan ole kuollut.

List 15

Mahdollisuudet lisääntyvät.
Poliiseja ei valmistu tarpeeksi.
Tavaroita kerätään kiireellä.
Naisten kilpailussa kärki erottui.
Teksti etenee puheen rytmissä.
Kone kotiin, ja sehän toimii.
Säilyykö palvelujen taso?
Ilo ja suru täydentävät toisiaan.
Ahosen ongelma on tekniikassa.
Vaihtoon Suomi tuli kakkosena.
Kummallakin on kymmenen osumaa.
Mitähän äiti ja isä tuumaavat?
Puistossa voi nauttia myös näytelmistä.
Maan pinnalla sen sijaan on vilkasta.
Halukkaita ei ole ilmaantunut.
Yhtiö rahoittaa norjalaisia kuntia.

List 16

Palvelu on käyttäjälle maksuton.
Sama historia on Rehnillä.
Valmista tietä on jo Paimioon saakka.
Vain puolet ryhmästä pääsi perille.
Mäenpää esiintyy itsekin juhlassa.
Miten joku voi tappaa itsensä?
Puhuessaan kuolleet todistavat.
Ammatti ei kätke tunteita.
Helsingissä on vastassa Belgia.
Mies myönsi kirjoittaneensa kirjeet.
Miesten kuulusteluja jatketaan.
Ihan mukavaltahan se kuulostaa.
Esitykset jatkuvat syksyllä.
Kiinnostaako presidentin vaalit?
Mikkeliin on mukava palata.
Silloin ottelu oli jo ratkennut.

List 17

Miten uudistus on toteutunut?
Jokerit kuittasi jo hetken päästä.
Vaara kilpaili Postin joukkueessa.
Asetus on parhaillaan lausunnolla.
Oma itsensä kannattaa olla.
Vierailu oli mielenkiintoinen.
Joskus idea lähtee materiaalista.
Viinanen ei jää tyhjän päälle.
Eniten vikaa löytyi jarruista.
Venäjälle Suomi hävisi.
Poika menehtyi onnettomuudessa.
Suomalainenkin kuulemma käy.
Miehetkin saavat liittyä yhdistykseen.
Tavanomainen ooppera ei riitä.
Ratkaisuun vaikuttaa salkun sisältö.
Ostajalla täytyy olla intressi.

List 18

Soittajalla on mestarin otteet.
Loppua kohden pidot parani.
Sopimus tuli voimaan kesällä.
Kiitos, kiitos, vastasi Kankkunen.
Asia kuitenkin elää mielessäni.
Turun ympäristössä pyöritään.
Tutkija kaipaa mainetta ja kunniaa.
Maksunsa saatuaan taksi ajoi pois.
Lasten päivä osuu Vanhusten viikkoon.
Isku oli todella kova.
Yrittäjiä ja veneitä on liikaa.
Viimeinen kierros oli täyttä tuskaa.
Hehän ovat kisan ainoat miehet.
Yleisöä alkaa valua radan varteen.
Ihmisen ja puun raja katoaa.
Jokainen ymmärtää leikin hengen.

List 19

Omia ennätyksiä on kiva rikkoa.
Asiat liittyvät tiiviisti toisiinsa.
Mikään ei kuitenkaan auttanut.
Ruotsalaiset valittavat helpommin.
Näyttely avataan perjantaina.
Terhillä menee tosi lujaa.
Esillä on myös eksoottisia ruokia.
Koulun historia julkaistaan syksyllä.
Samaan aikaan kustannukset kasvavat.
Hiihtäjien paikat ovat vielä jaossa.
Ilon määrää ei voi mitata.
Anuun pätee sama kuin Jereenkin.
Joukosta oli poissa Jugoslavia.
Ehkä se on lapsuuden perua.
Hinta näkyy kulutuksessa.
Edelleen haussa on pari miestä.

List 20

Oikeuteenkin asian voi viedä.
Lehtoselle kävi pahasti.
Poikkeukset vahvistavat säännön.
Selvittämistä jatketaan edelleen.
Yhtiö uskoo tulevaisuuteen.
Mukana on myös Suomen Kuntaliitto.
Apua on luvassa teknologiasta.
Ahtisaari on kärsinyt painostaan.
Mies oli jonkin verran humalassa.
Pelin loppu oli dramaattinen.
Useimmiten tosin sakot riittävät.
Heitähän hallituksella riittääkin.
Laaksonen luottaa valmennettaviinsa.
Pieni voi todellakin olla suurta.
Espanjan täytyy puolustaa rajojaan.
Markkoja yritys ei julkista.