CHAPTER 21

Quality Evaluation of Synthesized Speech

Vincent J. van Heuven

Dept. Linguistics/Phonetics Laboratory, Leiden University
P.O. Box 9515, 2300 RA Leiden, The Netherlands

Renée van Bezooijen

Dept. General Linguistics and Dialectology
University of Nijmegen
P.O. Box 9103, 6500 HD Nijmegen, The Netherlands

Contents

Speech Coding and Synthesis
Edited by W.B. Kleijn and K.K. Paliwal

1. Introduction

1.1. Speech coding versus speech synthesis

By speech synthesis we will mean a system that takes (the ascii representation of some) conventionally spelled unrestricted text and converts this to speech, i.e. a reading machine, alternatively called a text-to-speech system or TTS-system. From a quality assessment viewpoint, a TTS-system is more complex than a speech coder. In speech coding longer stretches of human input speech are encoded at a low bit rate, transmitted or stored, and decoded at the receiver with greater or lesser degradation due to information loss. Generally the ASSESSMENT of speech coding involves the direct quality comparison between the original human input speech and the output of the encoding/decoding process. TTS-output differs from speech coding output in at least two important respects. First, TTS-speech is generated by recomposing words and sentences from a finite set of synthesis building blocks (such as phonemes, diphones, demi-syllables, or some more flexible unit). The problem of incorrect transitions between successive units does not arise in the case of speech coding, but looms large in TTS- applications. Secondly, the adequacy of the speaker's (oral reading) performance is not under evaluation in speech coding. In TTS evaluation, however, we are not only dealing with the potential loss of sound quality due to some information reduction scheme, but also with assessing the quality of the oral reading performance of the machine: does it adequately express the intentions of the writer of the text in terms of its choice of words, speech melody and timing? Though there is an obvious partial overlap between evaluating coding schemes and TTS-systems, the differences between the two necessitate rather disparate evaluation techniques. This chapter aims to present a survey of current TTS evaluation practice.

1.2. Why speech synthesis evaluation?

In spite of the rapid progress that is being made in the field of speech technology, any speech synthesis system available today can still be spotted for what it is: nonhuman, a machine. Most older systems will fall through immediately due to their robot-like melody and garbled vowels and consonants. Other, more recently developed synthesis techniques using short-segment waveform concatenation techniques such as PSOLA [47] yield a segmental quality that is very close to human speech [59], but still suffer from noticeable defects in matters of melody and timing. As long as synthetic speech is inferior to human speech, synthesis evaluation will be useful. Speech synthesis assessment can be important to two parties: systems designers on the one hand, and prospective buyers and end users on the other. Designers are intent on improving their TTS-systems. However, designers who have grown up with their system are used to all its habits; they are likely to understand its output better than first-time users, and will often overrate its performance level. More meaningful quality assessment techniques are needed in order to determine

how well a system performs relative to a benchmark test, or how favorably it compares with a previous edition of the system or with an other designer's product. To the extent that a system performs less than perfect, the designer will have to learn which aspect(s) and/or component(s) of his system are flawed. Designers will therefore also be interested in diagnostic testing, either by doing detailed error analyses on the test results, or by running component- specific tests.

The needs of buyers and end users are different than those of designers but they, too, heavily rely on assessment techniques. Prospective buyers will always have a specific use of their TTS-system in mind. Understandably, they will want the simplest, and therefore cheapest, system that satisfies their needs. The buyer (or his consumer organization) will therefore need an absolute yardstick in order to determine beforehand if the TTS-system is good enough to get the message across in the given application. Buyers will not normally be interested in diagnostic testing.

1.3. History of synthesis evaluation

The history of speech synthesis evaluation cannot be older than the existence of speech synthesis itself. Although a number of attempts at constructing talking machines have been made through the centuries, such as the talking head by Albertus Magnus, the speaking machine by Wolfgang von Kempelen, and the hand-operated voder by Homer Dudley (for an overview cf. [18]), the quality of these systems was so appalling that formal evaluation procedures were never even considered.

It seems fair to say that output evaluation has been an integral part of the development of TTS-systems, ever since TTS was considered a serious application. The earliest TTS-system was developed at the Haskins Laboratories as a reading machine for the blind [51], and its formal evaluation was published only a year later [52], using test methodologies that were adopted mainly from audiology, i.e. developed to establish the extent of a patient's hearing loss. Audiological tests (such as the Harvard Psychoacoustic Sentences) yield adequate measures of segmental intelligibility, no matter whether the loss of quality resides with the speech producing apparatus (as in TTS) or in the listener (as is the case in hearing loss). The early audiological tests were not developed for diagnostic purposes; they established the amount of noise or signal distortion that a listener could bear before more than 50% of the words or syllables in a set of sentences could no longer be recognized. Obviously, if one wants to analyze the confusion patterns in the error responses for diagnostic purposes (see below), the test materials have to be constructed with this specific purpose in mind. Moreover, it soon transpired that the quality of TTS-systems could not be adequately tested without including such matters as rhythm and intonation. The audiological tests did not test rhythm and intonation perception, simply because these prosodic characteristics of human speech are not affected by hearing loss. As a result, TTS output testing methods were developed which differed from audiology tests.

1.4. Towards a taxonomy of evaluation tasks and techniques

To structure our overview of TTS assessment tests we will discuss a number of useful distinguishing parameters, which partly overlap with earlier attempted taxonomies (see e.g. [75, 58, 32]) and explain the relationships between them, before dealing with any specific assessment techniques.

The diagram shown in fig. 1 illustrates the relationships between the various dichotomies that make up our taxonomy. We will now discuss these six dichotomies in the hierarchical order in which they have been listed in this diagram.

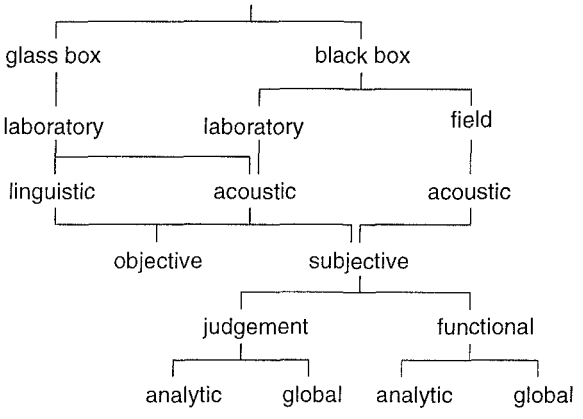

Figure 1. Relationships among dimensions involved in a taxonomy of speech output evaluation methods. Any path from the root down to any terminal that does not cross a horizontal gap constitutes a meaningful combination of test attributes.

1.4.1. Black box (monolithic) versus glass box (modular)

TTS-systems generally comprise a range of modules that take care of specific tasks. The first module converts orthographic input to some abstract linguistic code that is explicit in its representation of sounds and prosodic markers. Various modules then act upon this symbolic representation. Typically, one module concatenates the primitive building blocks (phonemes, diphones) in their appropriate order, another implements what coarticulation is needed to obtain smooth human-like transitions between successive building blocks. Prosodic modules, taking the positions of word stresses, sentence accents, phrasal and sentence boundaries into account, then provide an appropriate temporal organization (local accelerations and decelerations, pauses) and speech melody.

End users will typically be interested in the performance of a system as a whole. They will consider the system as a *black box* that accepts text and outputs speech, a monolith without any internal structure. For them it is only the quality of the output

speech that matters. In this way systems developed by different manufacturers can be compared or the improvement of one system relative to an earlier edition can be traced over time (*comparative testing*). However, if the output is less than optimal it will not be possible to pinpoint the exact module or modules that caused the problem. For *diagnostic* purposes, therefore, designers often set up their evaluations in a more experimental ("*glass box*") way. Keeping the effects of all modules but one constant, and systematically varying the characteristics of the latter, any difference in the assessment of the system's output must be caused by the variations in the target module. Modular testing, of course, presupposes that the researcher has control over the input and output of each individual module.

1.4.2. Laboratory versus field

TTS-systems are often part of a larger human-machine interface in a specific application. Typically, the vocabulary and types of information exchanges are restricted and domain-specific, so that situational redundancy is likely to make up for poor intelligibility. On the other hand, TTS-systems will often be used in complex information processing tasks, so that the listener has only limited resources available for attending to the speech input. Also, end users in the field may have different attitudes towards, and motivations for, working with artificial speech than subjects in laboratory experiments. It is generally impossible, therefore, to predict beforehand, on the basis of *laboratory tests*, exactly how successful a TTS-system will be in the practical application. The system will have to be tested in the field, i.e. in the real situation, with the real users. The use of *field tests* will be limited to one system in one specific application; results of a field test cannot, as a rule, be generalized to other systems and/or other applications.

1.4.3. Linguistic versus acoustic

The more complex TTS-systems can roughly be divided into a linguistic interface that transforms spelling into an abstract phonological code, and an acoustical interface that transduces this symbolic representation to an audible waveform. The quality of the intermediary representation can be tested directly at the *symbolic-linguistic* level or indirectly at the level of the *acoustic* output. Testing the audio output has the advantage that only errors in the symbolic representation that have consequences for the audio output, will affect the evaluation. The disadvantage of audio testing is that it involves the use of human listeners, and is therefore costly and time-consuming. Moreover, the results of acoustic testing are unspecific in that the designer is not informed whether the problems originate at the linguistic or at the acoustic level. As an alternative the intermediate representations in the linguistic interface are often evaluated at the symbolic level. It is, of course, a relatively easy task to compare the symbolic output of a linguistic module with some pre-stored key or model representation and determine the discrepancies, and this is what is normally done. The nontrivial problem is where to obtain the model representations. These will generally have to be compiled manually (or semi-automatically at

best), and often involve multiple correct solutions.

1.4.4. Subjective versus objective measurement

When an assessment technique involves the responses of human subjects, the measurement is called *subjective*. In a vast majority of cases human subjects are called upon in order to determine the quality of a TTS-system. This should come as no surprise to us, since the end user of a TTS-system is a human listener. However, there are certain drawbacks inherent to the use of human subjects. Firstly, humans, whether acting as single individuals or collectively as a group, are always somewhat noisy in their judgments or task performance, i.e. the results of tests involving human responses are never perfectly reproduceable. It often makes good sense to engage an expert listener as a short-cut to a preliminary evaluation. A professionally trained phonetician who is also a native speaker of the language, will generally be able to determine with great accuracy which vowels and consonants, and combinations thereof, are off the mark, and explain in articulatory terms what should be done to get the output right. To a lesser extent, the same can be done with temporal organization and intonation (cf. [68]). We would advocate such evaluations as a diagnostic tool in the initial stages of the development of a system. However, the phonetically trained listener will not be able to predict in numerical terms how well the TTS-system would perform as a communication tool with naive listeners. Obviously, if this is what we want to assess, we must turn to nonexpert listeners. In such cases, the human measurement instrument can be made less noisy if we do not engage a single listener but a group of listeners, and average responses over the larger group (which is sometimes called *intersubjective* measurement).

In addition to yielding noisy data, tests involving human subjects are time-consuming and therefore expensive to run. Recent developments seek to replace human evaluation by automatic quality assessment of TTS-systems, or modules thereof, automatically measuring the discrepancy in acoustical terms between a system's output and its human model. This is the type of *objective* evaluation technique that one would ultimately want to come up with: the use of human listeners is avoided, so that perfectly reproducible noiseless results can be obtained in as little time as it takes a computer to execute the program. At the same time, however, it will be clear that implementation of such techniques as a substitute for human listeners presupposes that we know exactly how human listeners evaluate differences between two realizations of the same linguistic message. Unfortunately, this type of knowledge is largely lacking at the moment and filling this gap may be difficult.

1.4.5. Judgment versus functional

By *judgment* testing we mean a procedure whereby a group of listeners is asked to judge the performance of a TTS-system along a number of rating scales. The scales are typically bi-polar adjectives that allow the listeners to express the quality of the system along a more global or more specific aspect of its performance.

Next, a TTS-system can be assessed in terms of how well it actually performs its

communicative purpose. This is called *functional* testing. For instance, if we want to know to what extent the output speech is intelligible, we may prefer to measure its intelligibility not by asking a listener how intelligible he *thinks* the speech is, but by determining, for instance, whether the listener correctly identifies the sounds.

1.4.6. Global versus analytic

Judgment tests usually include one or more rating scales covering such *global* aspects as "overall quality", "naturalness" and "acceptability". A functional approach to global assessment would be, for instance, to determine whether users of a TTS-system, when given the choice, choose to work with a machine or with the human original the machine is intended to simulate.

On the other hand, one may be interested in determining the quality of specific aspects of a TTS-system, in an *analytic* listening mode, where listeners are requested to pay particular attention to selected aspects of the speech output. Again, both judgment and functional tests can be, and have been, designed addressing the quality of specific speech aspects. Listeners may be asked to rate the clarity of vowels and consonants, the appropriateness of stresses and accents, pleasantness of voice quality, and tempo. Functional tests have been designed to test the intelligibility of individual sounds (e.g. phoneme monitoring), of combinations of sounds (e.g. syllable monitoring), of whole words (word monitoring) in isolation as well as in various types of context (e.g. [50, 60]).

2. Evaluation of linguistic aspects

2.1. Preprocessing

The first stage of a linguistic interface expands abbreviations, acronyms, numbers, special symbols, etc. to full-blown orthographic strings, and makes decisions on what to do with punctuation marks and other nonalphabetic symbols (e.g. parentheses).

There are no standardized tests for determining the adequacy of text preprocessors. Yet, even a superficial comparison of the few evaluation studies that are available on preprocessing reveal completely different sets of error categories (cf. [39, 40] on the evaluation of the CSTR (Centre for Speech Technology Research, Edinburgh) text preprocessor, and [79] on a text preprocessor for Dutch). What is clearly needed for the evaluation of text preprocessors, is a principled analysis of the various tasks a text preprocessor has to perform, focusing on those classes of difficulties that crop up in any (European) language. Procedures should be devised that automatically extract representative items from large collections of recent text (newspapers) in each of the relevant error categories, so that multi- lingual tests can be set up efficiently. Once the test materials have been selected, the correct solutions to, e.g., expansion problems can be extracted from existing databases, or when missing there, will have to be entered manually.

2.2. Grapheme-phoneme conversion

By grapheme-phoneme conversion we mean a module that accepts a full-blown orthographic input (i.e. the output of a preprocessor), and outputs a string of phonemes. The output string does not yet contain (word) stress marks, (sentence) accent positions, and boundaries. Since the correct phonemic representation of a normally spelled word depends on its linear context and hierarchical position within the linguistic structure (assimilation to adjacent words, stress shift, cf. chapter 17) the adequacy of grapheme-phoneme conversion modules should not, in principle, be tested on the basis of isolated word pronunciation (citation forms). In practice, however, this is precisely what is done. The reasons for this are threefold: (1) for many languages pronunciation databases (or machine readable pronouncing dictionaries) are available, which are exclusively based on isolated words, whereas (2) machine readable phonemic transcriptions of continuous prose are scarce, and (3) the adaptation rules for word pronunciation in context are assumed to be straightforward, exceptionless, and easy to implement. However, many of the adaptations are style and context dependent. Listener preferences have hardly been researched in this area (but cf. [35]).

The output of grapheme-phoneme converters is generally matched against a prestored list of correct transcriptions, which may or may not contain alternative pronunciations for a given word. The approach typically adopted is to equally weigh every single discrepancy between the system's proposal and the prestored model (in terms of omissions, additions or substitutions of phonemes). Such counts seem to adequately differentiate between grapheme-phoneme converters (cf. e.g. [58, 49]), but more sophisticated approaches may be considered that weigh the discrepancies between proposed and prestored transcription according to some perceptually relevant distance metric (cf. [12]).

2.3. Morphological decomposition

In morphological decomposition orthographic words are analyzed into morphemes, i.e. elements belonging to the finite set of smallest sub-word parts with an identifiable meaning. Morphological decomposition is necessary when the language/spelling allows words to be strung together without intervening spaces or hyphens so as to form an indefinitely large number of complex, longer words, such as in Dutch and German [1]. For many languages word-internal morpheme boundaries are referred to by the grapheme-phoneme conversion rules. For instance, the English letter sequence *sh* is pronounced as /S/ when it occurs morpheme internally as in *bishop*, but is pronounced as /s/ followed by /h/ when a morpheme boundary intervenes, as in *mishap*. Morphological decomposition is a notoriously difficult task,

[1] As an example of excessive compounding consider the (probably apocryphal) German *Reichseisenbahnenknotenpunktenhinundherschieber* 'State railways points man' or *Donaudampfschiffgesellschaftfahrts- kapita"n* 'captain of a steam ship for tourist trips on the river Danube'

as one input string can often be analyzed in a large number of different ways. The hard problem is choosing the correct solution out of the many possible solutions. An amusing example is the Dutch compound *belangstellende* 'interested person', for which the decomposition program suggested *bel+angst+ellende* 'misery due to fear of making phone calls', with deviating phonemes and stress pattern. This sort of ambiguity can only be solved by taking world knowledge into account [2].

As far as we have been able to ascertain, there are no established test procedures for evaluating the performance of morphological decomposition modules. Laver [39] tested the morphological decomposition module of the CSTR TTS-system on 500 words randomly sampled from a 85,000 word type list, which was compiled from a large text corpus and two machine-readable dictionaries. The output of the module was examined by hand, and proved correct at 70% (which seems rather low considering the fact that the elements of English compounds are generally separated by spaces or hyphens).

The Dutch morphological decomposition module MORPA (MORphological PArser, cf. [23]) compared the module's output with pre-stored morphological decompositions in a lexical database. In this comparison only segmentation errors were counted, in a sample of 3,077 (simplex and complex) words taken from weekly newspapers. The results showed that in 3% of the input the whole word, or part of it, could not be matched with any entry in the MORPA morpheme lexicon. The frequency of this type of error depends on the coverage of the lexicon. Erroneous analyses were generated in another 1% of the input words. In all other cases the correct morphological segmentation was generated, either as the single correct solution (44%), or as the most likely solution in an ordered list of candidate segmentations (48%), or as one of the less probable candidate solutions (3%).

2.4. Word stress

Stressed syllables are generally pronounced with greater duration, greater loudness (in terms of acoustical intensity as well as pre-emphasis of higher frequencies), and greater articulatory precision (no consonant deletions, more peripheral vowel formant values). Moreover, when a word is presented in focus (i.e. as expressing important information to the listener), a prominence-lending fast pitch movement is executed on the stressed syllable of that word. In many (so-called quantity-sensitive stress) languages, including English and Dutch, the position of the stress varies from word to word. However, stress position in these languages is predictable to a large extent by rules that look at (1) the internal make- up of words (in terms of the lexical categories of their constituent morphemes and the hierarchical relationships between them), and (2) at the segment structure of the syllables making up the morphemes (cf. e.g. [36]). However, English (and Dutch) have a proportion of idiosyncratic words that do not comply with the proposed stress rules. Therefore the

[2] Stochastic models trained on large data sets can make good approximations of world knowledge, often performing as well as humans.

coverage of stress rule systems has to be evaluated, and errors have to be corrected by including the exceptions in a dictionary.

Tests of stress modules have been performed only on an ad hoc basis, either checking the output of the rules by hand (see [4] for Italian), or automatically (using the phonemic transcription field in lexical databases containing stress marks (see [38] for Dutch), which in turn had been checked by hand in some earlier stage of the database development) [3].

Finally, the correctness of stress-shift will have to be verified by hand. Lexical look-up will not do, since the stress-shift rule is triggered by the wider syntactic/phonological context in which the target word occurs, e.g. *the poker is red 'hot* versus *he held a 'red hot poker.*

2.5. Syntactic parsing

Syntactic analysis lays the groundwork for the derivation of the prosodic structure needed to insert phonological phrase boundaries (which block stress shifts) and intonation domain boundaries (which block assimilation rules, trigger preboundary lengthening, pause insertion, and boundary marking pitch movements). Syntactic structure also determines (in part) which words have to be accented. Finally, lexical category disambiguation is often a by-product of a syntactic parser.

Although the syntactic parser is an important module in any advanced TTS-system, we take the view that, in principle, its development and evaluation does not belong to the domain of TTS-systems. Syntactic parsing is much more a language engineering challenge, developed for automatic translation systems, grammar checking, and the like.

2.6. Sentence accent

Appropriate accentuation is necessary to direct the listener's attention to the important words in the sentence, as well as to prevent the listener from paying undue attention to words whose referents are already known to him. Inappropriate accentuation may lead to misunderstandings and processing delays (cf. [67]). For this reason most TTS-systems provide for accent placement rules, which can be evaluated at the symbolic and the acoustic levels. In [45, 46]) symbolic output of a sentence accent assignment algorithm applied to four English 250 word texts (transcripts of radio broadcasts) was tested. The algorithm generated primary and secondary accents, which were rated on a 4-point appropriateness scale by three expert judges. In [74] a Dutch accent assignment algorithm was tested at the symbolic as well

[3] English presents a special problem in the assignment of stress. The elements of English compounds are typically separated by spaces, so that each element is erroneously treated as a word by itself. Moreover, the stressing of compounds in English partly depends on the semantic relationship between the words that make up the compound, and in part on purely lexical factors. A comparison of English compound stress rules developed by linguists and decision rules automatically extracted from hand-labeled phonetic databases has been reported by [66].

as the acoustic levels (onlyone type of accent is postulated for Dutch) using 8 isolated sentences and 8 short newspaper texts. Two important points emerged from this study: (1) correlations between the symbolic and the acoustic evaluations were significant but rather low, which means that tests at the symbolic level are no adequate substitute for acoustic tests, and (2) ratings for isolated sentences were more favorable than for sentences in paragraphs, which means that paragraph testing is necessary if the speech output system has to produce connected text.

3. Evaluation of acoustic aspects

3.1. General methodology

3.1.1. Test procedures

Test procedures can vary with respect to subjects, stimuli, and response modality. Examples of *subject variables* affecting evaluation results are ear-training [76] and experience with synthetic speech, whether acquired through training with (e.g. [19, 63]) or without feedback [56, 7]. The learning effect has been found to manifest itself after only a few minutes of exposure. However, there are indications that learning depends on the type of synthesis used [34].

Having established that the type of subject has an effect on the intelligibility of synthetic speech, one may wonder what implications this has for the choice of subjects in specific tests. In principle, subjects should be selected who are representative of the (prospective) users. Synthesis integrated in a reading machine for the blind should be tested with visually handicapped. Synthesis to be used by the general public for incidental purposes should be tested with a wide variety of naive subjects, including dialect speakers. And synthesis for long-term use should be tested at different points in time: at the beginning and after different periods of familiarization with the synthetic speech. This approach is to be recommended not only because of (possible) differences in the perception of the speech output, but also because motivation is known to play an important role in the effort people are willing to spend learning to understand suboptimal speech. If people have a choice between human and synthetic speech, the synthetic speech will have to be good in order to be accepted. However, if people have no choice, e.g. the visually handicapped who will have no access to a daily newspaper unless through synthesis (or braille), synthesis will be accepted more easily.

Stimuli typically vary along the following parameters: length (monosyllabic, disyllabic, polysyllabic), linguistic level (word, sentence, paragraph), open versus fixed stimulus set, meaningless (or rather lexically unpredictable) versus meaningful, phonetically balanced (in accordance with the statistical distribution of the phonemes in the language) or equal representation of each phoneme.

As for *response modality*, a distinction can be made between e.g.:

off-line (i.e. allowing time to think) identification tests using a closed set of response categories or an open mode, combined with spelling (leading to problems

in the interpretations of the responses) or unambiguous notation (placing the burden upon the subjects) (e.g. [52, 82, 5, 30]),
- on-line (i.e. requiring immediate response) identification tests, requiring the subject to decide whether the stimulus is a meaningless or meaningful word (the so-called lexical decision task) (e.g. [56]),
- off-line comprehension tests in which content questions have to be answered in an open or closed response mode (e.g. [57]),
- on-line comprehension tests requiring the subject to indicate whether a statement is true or not (the so-called sentence verification task) (e.g. [44]), and
- judgment tasks (always on-line) involving the rating of scales (e.g. [13, 29]).

3.1.2. Benchmarks

By a benchmark test we mean an efficient, easily administered test, or set of tests, that can be used to express the performance of a TTS-system (or some module thereof) in numerical terms. The benchmark itself is the value that characterizes some reference system, against which a newly developed system is (implicitly) set off. The benchmark is preferably chosen such that it represents a performance level that is known to guarantee reasonable user satisfaction. Consequently, if the performance of a new product exceeds the benchmark, its designer or prospective buyer is assured of at least a satisfactory product, and probably even better. Obviously, testing against a benchmark is more efficient than pairwise or multiple testing of competing products. At this time it is too early to talk about either existing benchmarks or benchmark tests. It is clear, however, that the development of benchmarking deserves high priority in the TTS assessment field.

3.1.3. Reference conditions

Next to a widely accepted benchmark, it would seem that designers of speech output systems should want to know how well their systems perform relative to some optimum, and what performance could be expected of a system that contains no intelligence at all. In other words, the designer is looking for topline and baseline reference conditions. As for the assessment of *segmental* quality, the following would seem adequate:

- The topline segmental reference condition will be some form of human speech produced by a designated talker, i.e. the same individual on whose speech the table values and synthesis rules were based, or who, in the case of concatenative synthesis, provided the basic synthesis units. The absolute topline reference will then be based on CD-quality digital speech. However, if the synthesis is parametric, the human reference speech, in an additional condition, should be analyzed and (re-)synthesized using exactly the same coding scheme that is employed in the speech output system to be tested [4]. Comparison of the synthesis with both

[4] This requirement can generally be fulfilled when LPC synthesis schemes are used. However, for a range of synthesizers (e.g. the Klatt and the JSRU synthesizers) no automatic parameter

the parametrized (coded) and the CD-quality top-line reference allows the researcher to determine whether further improvements can still be made in the synthesis system itself, or whether the synthesis is optimal within the limitations of the coding system adopted.

- A useful baseline in allophone synthesis would be one in which all segments retain their table values and are strung together merely by smoothing spectral discontinuities at segment boundaries. In the case of concatenative synthesis one could string together coarticulatory neutral phones (i.e. stressed vowels spoken between two /s/-es, or stressed consonants preceded by schwa and followed by an unrounded central vowel, cf. the 'neutrone' condition in [76]). Again, minimal smoothing can be applied to avoid spectral jumps.
- Recently, attempts have been made at creating a continuum of reference conditions by taking high-quality human speech and applying some calibrated distortion to it, such as multiplicative white noise at various signal-to-noise ratio's ('Modulated Noise Reference Unit or MNRU, cf. ITU-T Recommendation P.81), or time-frequency warping (TFW, ITU-T Recommendation P.85, cf. [9]; or T-reference, cf. [11]). Moreover, the perceived quality of TTS-systems has been shown to interact with the sound pressure level at which the speech output is presented, so that optimal SPL's have to be determined for each TTS-system separately before comparisons can be made. [17] shows that the MNRU in not suitable for the evaluation of synthetic speech. TFW of natural speech, however, provided a highly sensitive reference grid within which TTS-systems could be clearly differentiated from each other in terms of judged listening effort [33].
  The need for suitable topline and baseline reference conditions has clearly been recognized in the field of *prosody* testing.
- As a realistic topline, we advocate copying the temporal structures and speech melodies of a single designated professional human speaker onto the synthetic speech output.
- The optimal baseline for temporal structure would be a condition in which the smallest synthesis building blocks retain their original, unmanipulated durations as they were copied from the human original from which they were extracted (or, in the case of allophone synthesis, the phoneme duration table values, cf. [10]). This baseline condition, then, contains no intelligence, so that any improvement in the target conditions with duration rules must be due to the added explicit knowledge on duration structure. A reference in which segment durations vary at random (within realistic bounds) can be included for validation purposes, as an example of a 'very bad system'.
- As for testing speech melody, we most frequently find that the baseline condition is synthesized on a monotone, at a pitch level that coincides with the average pitch of the test items. This choice is rather arbitrary, however. In analogy with the random duration reference, a random melodic reference can be included for the sake of validation, by making the pitch go up and down within (physiologically

estimation is possible. The optimal parametric representation of human reference materials will then have to be found by trial and error, or the attempt should be abandoned.

and linguistically) reasonable limits.

In the area of *voice quality*, the problem of reference conditions has not been recognized. Generally, there seems to be little point in laying down a baseline reference for voice quality. The choice of a suitable topline would depend on the application of the speech output system. If the goal is personalized speech output (for the vocally handicapped) or automatic speaker conversion (as in interpreting telephony), the obvious topline is the speaker who is being modelled by the system, using the same coding scheme when applicable. When a general purpose (i.e. nonpersonalized) speech output system is the goal, one would first need to know the desired voice quality, i.e. ideal voices should be defined for specific applications, and speakers should be located who adequately represent the ideal voices.

3.2. Aspects of speech to be evaluated

Traditionally in phonetics (e.g. [1]) three layers are distinguished in speech: a segmental layer (related to short-term fluctuations in the speech signal, i.e. roughly within a time-window the length of a demi-syllable), a voice dynamics or prosodic layer (medium-term fluctuations, i.e. a domain of variable length, between a syllable and an Intonational Phrase), and a voice quality layer (long-term fluctuations). We will make the same distinction in the evaluation of acoustic aspects of TTS-systems (and have done so in the preceding sections as well), 3.2.1 being concerned with testing segments, 3.2.2 with prosody, and 3.2.3 with voice quality. Tests which relate to the complete TTS-output, in which all three layers are integrated, will be discussed in 3.2.4, and tests which explicitly take application aspects into consideration will be dealt with in 3.2.5. Finally, in 3.2.6 relationships among tests will be examined.

3.2.1. Segments

3.2.1.1. Functions The primary function of segments, i.e. the consonants and vowels in the language, is simply to enable listeners to recognize words. Generally, when the segments are sufficiently identifiable, words can be recognized regardless of the durations of the segments and the melodic pattern. In the experience of most researchers good quality (readily identifiable) vowels are afforded by even the simplest speech synthesis systems. One reason is that most coding schemes allow adequate parametrization of vocalic sounds (narrow band formants slowly varying with time). The synthesis of good quality consonants is an altogether different matter (due to multiple excitation signals, notion of formant not always applicable, abrupt spectral changes), and this is where most (parametric) synthesizers show defects. Moreover, since speech extends along the time dimension, segments early in the word in practice contribute more to auditory word recognition than later segments. Trailing segments, especially in long (i.e. poly- syllabic) words are often not needed to distinguish the word from its competitors. Also, stressed syllables tend to contribute more to a word's identity than segments in unstressed syllables. For these reasons

it makes sense to break down the segmental quality of TTS-systems for vowels and consonants in various positions within monosyllabic and polysyllabic words (initial, medial, final), and in stressed versus unstressed syllables.

3.2.1.2. Tests Compared to prosody and voice quality, the evaluation of the segmental aspect of synthetic speech has received most attention till now, (1) because good segmental quality is considered to be the main prerequisite for good overall quality, (2) because there is general agreement on the relevant categories in terms of which quality can be assessed, namely phonemes, and (3) because it is easy to establish. Near perfect segmental quality is essential for applications with a strong emphasis on the transmission of low-predictability information to untrained listeners, for example traffic information and reverse telephone directory services. In applications like these, where prosody can be minimally implemented, the required intelligibility level can be attained e.g. by making use of canned speech or concatenative, nonparametric synthesis. In other applications, where text-to-speech is preferred, it may perhaps not be necessary for each sound to be identified correctly. However, since very little is known as yet on the specific contributions of single sounds to overall intelligibility, synthesis designers have usually taken the pragmatic position that in principle all sounds should be identifiable. In that case detailed diagnostic testing of segmental quality remains to be defined.

3.2.1.2.1. Word level First considering segmental evaluation at the word level, it can be observed that most tests are functional, quality being expressed in terms of correct phoneme identification, modular, which means that other aspects of speech are kept constant or their influence reduced, and analytic, the attention of the listeners being explicitly directed at segments. Examples of functional, modular, analytic tests used to evaluate segmental quality of synthetic speech at the word level are the Diagnostic Rhyme Test (DRT), the Modified Rhyme Test (MRT), the Bellcore Test, the Cluster IDentification (CLID) Test, and the Minimal Pairs Intelligibility (MPI) Test.

3.2.1.2.2. DRT and MRT The DRT [82, 81] is a closed response test with two response alternatives containing systematic, minimal phonemic contrasts in the initial consonant. The subject would be asked e.g. to indicate whether a synthetic item was intended as *dune* or *tune*. The MRT [25] is an (originally) closed response test with six response alternatives differing either in the initial or the final consonant, e.g. *peas, peak, peal, peace, peach,* and *peat*. Both the DRT and MRT make use of meaningful words, which makes them reliable, fast, and easy to administer and score. No training is required of the subjects because the responses are in normal spelling. The tests are suitable instruments for comparative purposes at the word level. However, intelligibility may be overestimated since subjects adjust their perception to the response categories presented to them. Moreover, there is a risk

of ceiling effect. Finally, due to their restricted coverage and their limitation to meaningful words, the tests have little diagnostic value.

Both the DRT and MRT have been used extensively in TTS-evaluation. The DRT has been employed among others in [27], who compared a wide range of synthetic voices/systems and a human reference, both clear and with noise added to give a speech-to-noise ratio of 0 db(A). The percentages correct in the clear condition ranged between 61% and 96%. Adding noise extended the range to between 30% and 80%, making the test more sensitive. The MRT has been employed, among others, in [26] to evaluate eight synthesizers and a human reference. On the basis of the results, the systems were grouped into four categories, namely (1) human voice (99% correct, averaged over initial and final consonants), (2) high-quality TTS (95%), (3) moderate-quality TTS (85%), and (4) low-quality TTS (68%). The categories distinguished could be used as benchmarks (although somewhat dated, the set of synthesizers tested is probably representative of the quality range of more recent synthesizers).

3.2.1.2.3. Bellcore Test and CLID Test In the DRT and MRT no consonant clusters are included. The importance of this structure should not be underestimated. According to [65], about 40% of all one-syllable words in English begin and 60% end with consonant clusters. The Bellcore Test and the CLID Test have been developed to fill this gap. The CLID Test [30] is a very flexible architecture which can be used for generating a wide variety of monosyllabic stimuli (e.g. CCV, VCCC, CCCVVC) in an in principle unlimited number of languages as long as matrices are available with the phonotactic constraints to be taken into account. Both the intelligibility of (sequences of) initial and final consonants and of (sequences of) medial vowels can be tested.

In contrast to the CLID Test, the Bellcore Test [65] has a fixed set of stimuli, comprising both meaningless and meaningful words. Vowels are not tested, only (sequences of) consonants, which are tested separately in initial and final position. This makes the stimuli less complex and the task of the subjects less heavy. A disadvantage of the Bellcore Test is that no test material is available for other languages than English. The test has been applied to assess the intelligibility of two synthesizers compared with human speech, presented over the telephone [65]. The syllable score was 88% for human telephone speech and around 70% for the synthetic telephone speech.

3.2.1.2.4. MPI Test Finally, the Minimal Pairs Intelligibility Test (MPI Test, [80]), consists of a fixed set of 256 sentence pairs containing one contrast, e.g. *The horrid courts scorch a revolution* versus *The horrid courts score a revolution.* The minimal pair appears on the screen and the correct sentence has to be identified. The MPI Test was designed to expand the coverage of the DRT and MRT to include (1) vowels, (2) consonants in clusters, (3) unstressed syllables, (4) de-accented or cliticized words, (5) words in sentences, (6) polysyllabic words, and (7) insertions

and deletions. The test also aims at reducing ceiling effects, which arise since the DRT is not sensitive enough to differentiate between the better types of synthesis.

The MPI Test is a useful extension of the DRT/MRT paradigm, but at considerable cost. Although a wide range of diagnostic information is obtained, it is not done in an efficient way, since each response gives information on the identifiability of only one phoneme. Moreover, creating test materials presupposes the availability of large databases.

3.2.1.2.5. Judgment tests In principle, in addition to functional intelligibility tests, judgment tests, where subjects rate the stimuli on scales, are possible for evaluating the segmental quality at the word level as well. For example, [71], in addition to running a cluster identification test, presented 26 Dutch consonant clusters (both initial and final) to be rated on naturalness, intelligibility, and pleasantness. The clusters were embedded in meaningful words and subjects were explicitly asked to pay attention to the clusters only. So, the test required analytic listening. However, one can never be sure to what extent listeners in fact stick to the instructions. Perhaps this is one of the reasons why judgment tests of this type have been rare.

3.2.1.3. Sentence level Tests for the assessment of segmental quality have also been developed at the sentence level. Compared with the segmental tests at the word level, tests at the sentence level are more similar to speech perception in normal communication but at the same time, as a consequence, less suitable for diagnostic purposes. Firstly, with sentences, the intelligibility scores will not only be based on segmental quality but also to some extent on prosodic quality, so that poor intelligibility is more difficult to trace back to specific sources. Secondly, the composition of the test material is somewhat unsystematic, so that no complete confusion matrices can be obtained. Moreover, especially with semantically normal sentences listeners will not only rely on segmental information but use other information sources as well, related to word internal and word combinatory redundancy. Of course, if the test is not intended as a diagnostic tool but has a purely comparative aim, these consequences do not necessarily detract from its value.

In this section only functional tests will be discussed, namely the Harvard Psychoacoustic Sentences, the Haskins Syntactic Sentences, and the Semantically Unpredictable Sentences (SUS). In addition, judgment tests at the sentence level have frequently been carried out. These are described in 3.2.4 under *overall output quality*. They entail the rating of scales such as *acceptability, intelligibility,* and *naturalness.*

subparagraphHarvard Psychoacoustic Sentences and Haskins Syntactic Sentences

One of the most well-known intelligibility tests at the sentence level is the fixed set of 100 semantically and syntactically normal Harvard Psychoacoustic Sentences (*Add salt before you fry the egg*) [16]. The test is easy to administer (no training required of the subjects) and score (be it manually). However, there is a strong learning effect and a danger of ceiling effect.

Another famous test at the sentence level is the fixed set of 100 semantically unpredictable Haskins Syntactic Sentences (*The old farm cost the blood*) [52]. Just like the Harvard Sentences, the Haskins Sentences are easy to administer and score. But here also there is a learning effect, so that subjects can be used only once. Moreover, generalizability is limited, since there is only one syntactic structure. The Haskins sentences were applied to four synthesizers and human speech by [57], and compared with the Harvard sentences. The two tests yielded the same rankorder. However, as expected, the Haskins sentences were more sensitive.

3.2.1.3.1. Semantically Unpredictable Sentences More recently, a lexically open approach was opted for in the Semantically Unpredictable Sentences (SUS) developed by SAM (see [28], Chapter 5). The SUS test consists of a fixed set of five syntactic structures which are common in most Western European languages. The lexical slots are filled with high-frequency words from language specific lexica. Pilot studies have been run in French, German, and English [5, 6, 22].

3.2.2. Prosody

3.2.2.1. Functions By prosody we mean the ensemble of properties of speech utterances that cannot be derived in a straightforward fashion from the identity of the vowel and consonant phonemes that are strung together in the linguistic representation underlying the speech utterance. Prosody would then comprise the melody of the speech, word and phrase boundaries, (word) stress, (sentence) accent, tempo, and changes in speaking rate. We exclude from the realm of prosody the class of voice quality features (see 3.2.3).

Prosodic features may be used to differentiate between otherwise identical words in a language (e.g. *trusty* *trustee*, with initial stress versus final stress, respectively). Yet, word stress is not so much concerned with making lexical distinctions (this is what vowels and consonants are for) as with providing checks and bounds to the word recognition process. Hearing a stressed syllable in languages with (more or less) fixed stress informs the listener where a new word may begin; error responses in word recognition strongly tend to agree with the stimulus in terms of stress position. The more important functions of prosody, however, are located at the linguistic levels above the word:

- prosody offers segmentation cues in the form of phrase boundaries, i.e., it tells the listener which words go together and should be interpreted as making up a coherent chunk of information; also, these cues allow the listener to determine the "depth" of the break between chunks, i.e., whether he has come to the end of a word group, clause, sentence, or even a whole paragraph,
- prosody provides an indication for the listener which words are presented by the speaker as expressing important information (highlighting or focusing through accentuation),
- prosody, especially melody, carries some meaning of its own (intonational meaning) which, for example, allows the speaker to present a sentence as a statement

or a question, or to express his emotions and/or attitude towards the verbal contents of the message or towards the hearer.

These functions suggest that prosody affects comprehension rather than intelligibility and, indeed, comprehension is what most functional tests of prosody aim to evaluate.

3.2.2.2. Tests

3.2.2.2.1. Judgment evaluation Judgement evaluation of TTS-prosody is alternately focused on the formal or the functional aspects. Only a handful of tests are directed at the formal quality of *temporal organization*. An exemplary evaluation study on the duration rules of MITalk [3] was done by [10], using proper baseline and topline reference conditions as explained in section 3.1.3. Their results showed that the temporal organization afforded by the complete rule set was judged as natural as the human topline control. Moreover, sentences generated with boundary markers at minor and major breaks were judged more natural than speech without boundary markers [5]. More work has been done in the field of *melodic structure*. The *formal properties* of, for example, pitch movements or complete speech melodies can be tested by asking groups of listeners to state their preference in pairwise comparisons or to rate a melody in a more absolute way along some goodness or naturalness scale. At the level of elementary pitch movements (such as accent-lending or boundary marking rises, falls, or rise-fall combinations) the SAM Prosodic Form Test [20] is a useful tool.

Using the same methodology, i.e. rating and pairwise comparisons, the quality of synthetic speech melody can be evaluated at the higher linguistic levels. At the level of isolated sentences pairwise comparisons of competing intonation-by-rule modules is feasible when the number of systems (or versions) is limited (e.g. [2]). When multiple modules are tested using a larger variety of sentences and melodies, scale rating is to be preferred over pairwise comparisons for reasons of efficiency [15, 84]. Evaluation of speech melody generators should not stop at the level of isolated sentences. Ratings by expert listeners in Dutch could not reveal any quality differences between synthetic melodies and a human reference when the sentences were listened to in isolation; however, the same synthetic melodies proved inferior to the human reference when they were presented in the context of their full paragraph [69].

[5] Later (cf. [3]), the duration rules were evaluated directly (objectively) by comparing the predicted segment durations with the segment durations as measured in spectrograms of new paragraphs read by the designated speaker. The rules accounted for 84% of the duration variance with a residual standard deviation of 17 ms (excluding the prediction of pause duration). Seventeen ms is generally less than the just noticeable difference for a duration change in a single segment in a sentence context [37], which would explain why the human reference and the rule-derived durations were judged equally natural.

There is (at least) one judgment test that assesses how well certain communicative *functions* are signaled by prosody at a higher level. The SAM Prosodic Function [6]. Test [21] asks for ratings of the communicative appropriateness of melodies in the context of plausible human-machine dialogue situations. The test was applied to human- machine dialogues designed to simulate a telephone enquiry service giving flight information.

Finally, we are not aware of tests asking subjects to judge the quality of the expression of *emotions and attitudes* in synthetic speech. It would appear that functional testing of these qualities is preferred in all cases.

Evaluating TTS-prosody using *functional tests* is even more in its infancy. Since prosody is highly redundant given the segmental information (with the exception of the signaling of sentence type and emotion/attitude), it can be functionally tested only if measures are taken to reduce its redundancy. This is achieved by degrading the segmental quality, such that without prosody (i.e. in the baseline conditions identified above) the intelligibility of the TTS-output would be extremely poor. The quality of the prosody would then be measured in terms of the gain in intelligibility, i.e. increase in percent correctly reported linguistic units (phonemes, morphemes, words) due to the addition of prosody. [10] measured intelligibility of utterances synthesized by MITalk with and without application of vowel duration, consonant duration and boundary marking rules (see above). They found that adding duration rules improved word intelligibility; adding within-sentence boundaries, however, did not boost intelligibility (even though the result was judged to be more natural, see above). [62] demonstrate that adding within-sentence boundaries (i.e. changing the temporal organization) does improve word intelligibility (especially for monosyllabic words) in Dutch diphone synthesis, and that utterances with pauses were judged as more pleasant to listen to [78].

There is a substantial literature on the perception of emotion and attitude in human speech (for a survey, see [48]). Typically, listeners are asked to indicate which emotion they perceive in the stimulus utterance, in open or closed response format. Predictably, the larger the set of response alternatives, the poorer the identification of each emotion. Results tend to show that the most basic emotions can be identified, in lexically neutral utterances, at better than 50% correct, in a 10 alternative closed response test. Synthesis of emotion is being attempted by several research groups. Preliminary evaluation of emotion-by- rule in Dutch diphone synthesis was presented by [83].

3.2.3. Voice quality

3.2.3.1. Functions Whereas the segmental and prosodic features of speech are continuously varying, voice quality is taken to refer to aspects of speech which generally remain relatively constant over longer stretches of speech. Voice quality can be most

[6] The notion 'function test' in this sense has no relationship with our use of the term 'functional test'. In the SAM Prosodic Function Test prosodic quality is not being tested in a functional task: we are still dealing with intuitive judgments (ratings) of how well the melody would fulfil its function without actually testing it.

easily viewed as the background against which segmental and prosodic variation is produced and perceived. In our definition, it includes such varied aspects of speech as mean pitch level, mean loudness, mean tempo, harshness, creak, whisper, tongue body orientation, dialect, accent, etc. Voice quality is mainly used by the listener to form a (sometimes incorrect) idea of the speaker's mood and personality (cheerful, reliable, dominant), physical size (tall, large, strong), sex (male, female), age (child, young adult, aged), regional background (globally "from the North" or more precisely "from London, Paris, or New York"), socio-economic status (high/low education), health (cold), and also to identify the speaker. This information may have practical consequences for the continuation of the communicative interaction, since it may influence the listener's attitudes towards the speaker in a positive or negative sense and may affect his/her interpretation of the message (cf. [42]).

Since recently, increased attention is being paid to voice quality aspects of synthetic speech. In fact, [64] regards the successful creation of personalized synthetic voices ("personalized TTS") as one of the most ambitious challenges of the near future. This aspect of synthesis is, for example, relevant in such applications as Translating (Interpreting) Telephony services, where along with translating the content of the message the original voice of the speaker has to be reconstructed (automatic voice conversion). Moreover, the correct encoding of speaker characteristics such as sex, age, and regional background is also relevant for the reading of novels for the blind. Finally, a third application is to be found in nonspeaking disabled individuals, who have to use a synthetic speech to replace their own.

3.2.3.2. Tests Apart from specific requirements imposed by concrete applications, a general requirement of the voice quality of synthetic output is that it should not sound unacceptably unpleasant. *Voice pleasantness* is one of the scales included in the overall quality test proposed by the ITU-T to evaluate synthetic speech transmitted over the telephone. It has also been used by [73] in a field test to evaluate the functioning of an electronic newspaper for the blind. Interestingly, the pleasantness of voice ratings were found not to change over time, in contrast to the intelligibility ratings, which reflected a strong learning effect. From this it was concluded that voice quality has to be good right from the start; one cannot count on the beneficial effect of habituation.

Of course, judgment studies such as these can only provide global information; if results are negative, no diagnostic information is available as to what voice quality component should be improved. There are no standard tests to diagnostically evaluate the voice quality characteristics of TTS-output. This type of information could in principle be obtained by means of a modular test, where various acoustic parameters affecting voice quality are systematically varied so that their effect on the evaluation of voice quality can be assessed. This would be the most direct approach.

A more indirect approach would involve asking subjects to listen analytically to and rate various aspects of voice quality on separate scales. A potentially useful instrument for obtaining a very detailed description is the Vocal Profile Analysis

Protocol developed by [41]. This protocol, which comprises more than 30 voice quality features, requires extensive training. If data are available for several synthesis outputs the descriptive voice quality ratings could be used to predict the overall pleasantness of voice ratings.

It may also be possible to use untrained listeners, although the number of aspects described will necessarily be more limited and less "phonetic". Experience with human speech samples representing various voice quality settings [70] has shown that naive subjects can reliably describe 1-minute speech samples with respect to the following 14 voice quality scales: warm - sharp, smooth - rough, low - high, soft-loud, nasal - free of nasality, clear - dull, trembling - free of trembles, hoarse - free of hoarseness, full - thin, precise-slurred, fast-slow, accentuated - unaccentuated, expressive - flat, and fluent - halting. Again, if descriptive ratings of this type were available for synthetic speech they could be correlated with global ratings of synthesized voice quality. Alternatively, this type of scale could also be used more directly for diagnostic purposes, i.e. subjects could be asked to rate each of these voice quality aspects on a 10-point scale, with 1: extremely bad and 10: extremely good.

However, as mentioned above, experience with detailed perceptual descriptions of voice quality is as yet limited to nondistorted human speech. It remains to be assessed whether such descriptions can also be reliably made for synthetic speech. And even if this proved to be the case, the translation of the results obtained to actual system improvement is not unproblematic, since not much is known about the acoustic basis of perceptual voice quality ratings. Attempts in this direction have been rather disappointing (e.g. [8]).

In addition to judgment tests to evaluate the formal aspects of voice quality, functional tests may be used to assess the adequacy of voice quality. Although here also no standard tests are available, the procedures are rather straightforward and dictated directly by application requirements. One can think, for example, of tests in which subjects are asked, in an open or closed response format, to identify the speaker. This would be useful in an application where one tries to construct a synthetic voice for a given speaker or reconstruct the natural voice of a given speaker. Or one can ask people to identify the speaker's sex, or estimate his/her age or other characteristics.

3.2.4. Overall output quality

3.2.4.1. Preliminary remarks The functional quality of TTS-systems has mainly been evaluated by means of intelligibility tests in which listeners are required to "transcribe" sounds, resulting in a percentage correct identification of individual segments. The tasks performed in these laboratory tests, described in 3.2.1, resemble to some extent real-life situations where listeners have to identify unknown names of people or places. However, in most situations good intelligibility is not enough for TTS-output to be called functionally adequate. For general evaluation purposes, independent of the concrete aspects of contexts of application, one would want to have at one's disposal a functional test to evaluate the adequacy of the

complete TTS- output in all respects: does the output function as it should? Such a test does not exist, and is difficult to conceive. In practice, the *functional* quality of overall TTS-output has been equated with comprehension, based upon the integration of "bottom-up" speech signal information at different levels (segments, prosody, voice quality) and "top-down" knowledge and expectations based on previous experience, specific properties of the extra-linguistic context, and word internal and word combinatory redundancy.

3.2.4.2. Tests No completely developed standardized test, with fixed test material and fixed response categories, for evaluating comprehension is available, but one wonders whether this would be very useful in the first place, since at this level of evaluation it seems a good idea to take at least the content aspects of applications into account [7]. Testing the comprehensibility of TTS destined to provide traffic information asks for a more specific type of test materials than TTS to be used for reading a digital daily newspaper for the blind, where the test materials should cover a wide range of topics and styles. As to the type of comprehension test, several general approaches can be outlined. The most obvious one involves the presentation of synthesized texts at the paragraph level, preferably with human produced versions as a topline control, with a series of open or closed (multiple choice) questions.

At first sight, the results of closed response comprehension tests obtained in different studies seem to be somewhat counterintuitive: Although the human produced texts sound better than the synthetic version, often no difference in comprehension is revealed [53, 14] or, after a short period of familiarization, even superior performance for synthetic speech [56] is observed. These results have been tentatively explained by hypothesizing that subjects may make more of an effort to understand synthetic speech. Results of studies aimed at testing this hypothesis [44, 43, 7] are contradictory.

An example of an open response comprehension test is [72], who found significant differences among two synthesized and a human produced version of text passages. So, analogous to segmental intelligibility at the word level, an open response approach appears to be more sensitive than a closed response approach. However, the results also suggest that the effect of the supposedly greater effort expended in understanding synthetic speech has its limits. If the synthetic speech is bad enough, increased effort cannot compensate for loss of quality.

Other, more psycholinguistic approaches directly or indirectly related to comprehension have been developed and applied as well. To name but a few: (1) the word monitoring task, where subjects are instructed to press a button as soon as they hear a word out of a limited set of prespecified words, (2) the sentence-by-sentence listening task, in which subjects push a button whenever they are ready for hearing the next sentence (comprehension is checked afterwards but is not part of the

[7] Clearly, there is a continuum from completely application independent at the one end to completely application specific at the other end. The distinction between sections 3.2.4 and 3.2.5 is therefore somewhat artificial.

test proper), and (3) the sentence verification test, where subjects have to decide whether short sentences are true statements or not (e.g. *Mud is dirty* and *Rockets move slowly*). All three are on-line measures, the first indexing cognitive workload, the second and third assessing speed of comprehension. It has been suggested that tests of this type, which invariably reveal differences between human and synthetic speech, could be more sensitive than off-line measures, where responses are to be given after subjects heard test passages [60].

The approaches described so far to assess overall quality are functional in nature, employing a black box, global approach. A similar approach can also be used in *judgment tests* where subjects are asked to give their impression of global quality aspects of the TTS-output. Taking overall intelligibility as a concrete example, one can think of paired comparison tasks, where subjects indicate which of two synthesizers sounds more intelligible, magnitude estimation, where subjects assign a value or draw a line of a length which is equal to the magnitude of his impression of intelligibility, and categorical estimation, where subjects rate e.g. a 10- point scale which runs from 1: extremely unintelligible to 10: extremely intelligible. The magnitude estimation method is relatively laborious and more fit for test external comparison, whereas the categorical estimation method is relatively fast and easy, and more fit for test internal comparison.

Both the magnitude and categorical estimation methods have been included in the SAM Overall Quality Test (see [28], Chapter 7). Three scales are recommended, related to acceptability (the overall user's satisfaction with the communicative situation), intelligibility (how identifiable does the message sound), and naturalness (to what extent does the message sound like being produced by a human speaker). The intelligibility and naturalness ratings are based on pairs of unrelated sentences, to be synthesized with the TTS-system at hand. Fixed lists of 160 sentences of varying content and length are available for Dutch, English, French, German, Italian, and Swedish. An example for English is: *I realise you're having supply problems, but this is rather excessive.* For the acceptability ratings, application specific test materials are recommended.

The importance of application-specific test materials is also stressed by International Telecommunication Union Telecommunication Standardization (ITU-T) sector. They developed a test specifically aimed at evaluating the quality of telephone speech (where synthesis could be the input). It is a judgment test comprising ratings on (a subset of) eight scales, namely one 2-point scale *acceptance* and seven 5-point scales *overall impression, listening effort, comprehension problems, articulation, pronunciation, speaking rate,* and *voice pleasantness.* Strictly speaking, only the first four scales can be captured under the heading *overall quality*; the other four scales are directed at more specific aspects of the output and require analytic listening. The content of the speech samples presented should be in accordance with the application. In addition to rating the eight scales, subjects are required to reproduce information contained in the message. A pilot study has been run by [11].

3.2.5. Applications

3.2.5.1. Preliminary remarks As mentioned above, at the level of overall quality evaluation it is recommended to take application into account, for ultimately it is only the functioning of the TTS-output in an applied form that counts. If one is exclusively interested in comprehension, this will mainly have consequences for the content of the texts evaluated. Also, if only the test materials are adapted, the test can still be run in a laboratory setting. However, if overall quality is extended to include all aspects of the synthesis in an applied context, the consequences for evaluation may be more comprehensive and testing may be necessary in the field.

3.2.5.2. Tests A comprehensive field test, with equal attention to the TTS-output itself and the context within it is used, was done by [73]. They used a suite of four tests to evaluate the functioning of an electronic newspaper for the visually handicapped by means of (1) an interview enquiring after the subjects' attitudes towards the technology, (2) an open response identification tests with CVC-words, (3) judgments on 10 evaluative scales (1: extremely bad, 10: extremely good) related to global quality and more specific aspects of the TTS-system, such as pleasantness of voice (sec 3.2.3), adequacy of word stress, tempo, liveliness, and fluency, and (4) a functional test to evaluate the extent to which the subjects were able to find their way through the newspaper. The latter test involved a number of searches, such as *Is there an article on Japan in the economy section?* Performance was assessed both in terms of the number of correct answers and the time needed to accomplish the task. Comparable field tests have been conducted to evaluate a digital daily newspaper in Sweden [24]. And a similar combination of judgment and functional testing was done by [61] within the context of telephone information services.

3.2.6. Relationships among tests

Knowledge about the relationships among tests seems to us to be of great importance for two reasons: (1) it allows a better interpretation of the meaning and validity of the test results obtained, and (2) it can be used to decide upon the test suite which gives a complete picture of all relevant aspects of TTS without being redundant. It is no use to employ two tests which (to a large extent) yield the same information. Some differences between the results obtained with different tests can be predicted to some extent. For example, when considering intelligibility, we think at least four factors will affect the outcomes: Intelligibility can be expected to increase (1) as the unit of measurement is smaller (it is easier to identify one phoneme correctly than a sequence of phonemes), (2) as the structure of the test items is more predictable (fixed versus open structure), (3) as the combination of phonemes is more predictable (meaningful versus meaningless), and (4) as the number of response categories is smaller (closed versus open). These predictions can be tested by looking at actual intelligibility results. [31], for example, assessed the intelligibility of one German synthesizer by means of four different tests: the SAM Standard Segmental Test, a reduced version of the CLID test with single initial

and final consonants and three fixed vowels in medial position (open response), a German variant of the MRT (closed response), the CLID test (open response), and the SUS test (open response). Percentages correct elements (phonemes in the SAM Standard Segmental Test, clusters in the CLID test, words in the MRT and the SUS test) differed widely, from 19% to 85%. The lowest percentage was obtained for the SUS test, followed by the SAM Standard Segmental Test, the CLID test, and the MRT. The fact that the highest score was obtained with the MRT agrees with our predictions, since this test possesses not a single aspect with a negative effect on intelligibility: The unit of measurement is small (phoneme), the structure is fixed (CVC), the items are meaningful, and the response set is closed (six categories). The results for the other three tests point to complex interactions among the four factors.

4. Epilogue

In this chapter we have presented the state of the art of current speech synthesis testing. At this time, the development of speech synthesis seems to be branching off into two different directions. One the one hand, highly complex systems are under development, featuring excellent segmental quality due to waveform concatenation technology. In order to ascertain whether the segmental quality of these types of synthesis still falls short of natural human speech, highly sensitive assessment techniques are called for. We suggest that much can be learned from the related field of assessment of telecommunications systems [33] (see also chapter 13). For high-quality systems, further improvements will have to be sought mainly in the field of prosody, and in the quality of the linguistic modules that drive the prosodic rules. It seems to us that the development of diagnostic test techniques should be concentrated on these areas.

The other development in speech synthesis is the rise of low-budget technology, for use in the consumer's home, as aids for the visually handicapped, or as a human-machine interface for pre-school children and illiterate adults. Such cheap systems are often multi-lingual parametric synthesizers, that is, the hardware allows limited quality only, and the rules and exceptions dictionaries have been adapted, quick and dirty, to suit the needs of yet another language. Continued overall assessment of such systems remains necessary in order to insure that they are marketable at all; improvements of the systems are often feasible, and can be guided by the results of diagnostic testing, using the techniques outlined in this chapter.

Acknowledgment

Writing this chapter was made financially supported by EAGLES (Expert Advisory Group on Language Engineering Standards), an EC ESPRIT-III initiative. An expanded version of this chapter can be found in [77]. The present authors thank the following persons for comments on an earlier version of this chapter: Christian Benoit, Michel Cartier, Christina Delogu, Klaus Fellbaum, Adrian Fourcin, Valerie Hazan, Ute Jekosch, Denis Johnston, Louis Pols, Ann Syrdal, and one anonymous

reviewer.

References

[1] D. Abercrombie, *Elements of General Phone- tics.* Edinburgh University Press, Edinburgh, 1967.

[2] G. Akers and M. Lennig, "Intonation in Text-to-Speech Synthesis: Evaluation of Algorithms," J. Acoust. Soc. Am., Vol. 77, pp. 2157-2165, 1985.

[3] J. Allen, M.S. Hunnicutt, and D.H. Klatt, *From Text to Speech: The MITalk System.* Cambridge University Press, Cambridge, 1987.

[4] S. Barber, R. Carlson, P. Cosi, M.G. Di Benedetto, B. Granstro"m, and K. Vagges, "A Rule-Based Italian Text-to-Speech system," Proc. Eurospeech '89, Paris, Vol. 2, pp. 517-520, 1989.

[5] C. Benoit, "Intelligibility Test for the Assessment of French Synthesizers Using Semantically Unpredictable Sentences," Proc. of the ESCA Workshop on Speech Input/Output Assessment and Speech Databases, Noordwijkerhout, pp. 1.7.1 - 1.7.4, 1989.

[6] C. Benoit, A. Van Erp, M. Grice, V. Hazan, and U. Jekosch, "Multilingual Synthesizer Assessment Using Semantically Unpredic- table Sentences," Proc. Eurospeech '89, Paris, Vol. 2, pp. 633-636, 1989.

[7] T. Boogaart and K. Silverman, "Evaluating the Overall Comprehensibility of Speech Synthesizers," Proc. 2nd International Confe- rence on Spoken Language Processing, Banff, pp. 1207-1210, 1992.

[8] L. Boves, *The Phonetic Basis of Perceptual Ratings of Running Speech.* Foris, Dordrecht, 1984.

[9] M.D. Burrell, "Assessment of the Degradations of Synthetic Speech and Time Frequency Warping over Different Listening Levels," Proc. Institute of Acoustics, Vol. 13, Pt. 2, 1991.

[10] R. Carlson, B. Granstro"m, and D.H. Klatt, "Some Notes on the Perception of Temporal Patterns in Speech," Proc. 9th International Congress of Phonetics Sciences, Copenhagen, Vol. 2, pp. 260-267, 1979.

[11] M. Cartier, F. Emerald, D. Pascal, P. Combescure, and A. Soubigou, "Une Me'thode d'Evaluation Multicrite're de Sorties Vocales: Application au Test de 4 Syste'mes de Synthe'se a' Partir du Texte," 19e'mes Journe'es d'Etude sur la Parole, Bruxelles, 1992.

[12] C. Cucchiarini, *Phonetic Transcription: A Methodological and Empirical Study.* Doctoral Dissertation, University of Nijmegen, 1993.

[13] C. Delogu, A. Paolini, P. Pocci, and C. Sementina, "Quality Evaluations of Text-to-Speech Synthesizers Using Magnitude Estimation, Categorical Estimation, Pair Comparison and Reaction Time Methods," Proc. Eurospeech '91, Genova, pp. 353-355, 1991.

[14] C. Delogu, A. Paoloni, and C. Sementina, "Comprehension of Natural and Synthetic Speech: Preliminary Studies," ESPRIT SAM Final Report, II.e, 1992.

[15] J.R. De Pijper, *Modelling British English Intonation.* Foris, Dordrecht, 1983.

[16] J.P. Egan, "Articulation Testing Methods," Laryngoscope, Vol. 58, pp. 955-991, 1948.

[17] K. Fellbaum, H. Klaus, and J. Sotscheck "Horversuche zur Beurteilung der Sprachqualita"t von Sprachsynthesesystemen fu"r die Deutsche Sprache," in: *Fortschritte der Akustik. Plenarvortra"ge und Fachbeitra"ge der 20. Deutschen Jahrestagung fu"r Akustik*, DPG GmbH, Dresden, pp. 117-122, 1994.

[18] J.L. Flanagan, *Speech Analysis, Synthesis and Perception*, Springer, Berlin, 1972.

[19] S.L. Greenspan, H.C. Nusbaum, and D.B. Pisoni, "Perception of Speech Generated by Rule: Effects of Training and Attentio- nal Limitations," Research on Speech Perception Progress Report 11, Indiana University, pp. 263-287, 1985.

[20] M. Grice, K. Vagges, and D. Hirst, "Assessment of Intonation in Text-to-Speech Synthesis Systems - A pilot Test in English and Italian," Proc. Eurospeech '91, Genova, Vol. 2, pp. 879-882, 1991.

[21] M. Grice, K. Vagges, and D. Hirst, "Prosodic Form Tests," ESPRIT SAM Final Report Year Three, So.5, 1992.

[22] V. Hazan and M. Grice, "The Assessment of Synthetic Speech Intelligibility Using Semantically Unpredictable Sentences," Proc. of the ESCA Workshop on Speech Input/Output Assessment and Speech Databa- ses, Noordwijkerhout, pp. 1.6.1-1.6.4, 1989.

[23] J. Heemskerk and V.J. Van Heuven, "MORPA, a Morpheme Lexicon Based Morphological Parser," in: *Analysis and Synthesis of Speech, Strategic Research towards High-quality Text-to-Speech Generation*, Eds. V.J. Van Heuven and L.C.W. Pols, Mouton de Gruyter, Berlin, pp. 67-85, 1993.

[24] E. Hjelmquist, B. Jansson, and G. Torell, "Psycho- logical Aspects on Blind People's Reading of Radio-distributed Daily Newspapers," in: *Work with Display Units 86*, Eds. B. Knave and P.G. Wideba"ck, North-Holland, Elsevier Science Publishers, Amsterdam, pp. 187- 201, 1987.

[25] A.S. House, C.E. Williams, M.H.L. Hecker, and K.D. Kryter, "Articulation Testing Methods: Consonantal Differentiation with a Closed Response Set," J. Acoust. Soc. Am., Vol. 37, pp. 158-166, 1965.

[26] J.S. Logan, D.B. Pisoni, and B.G. Greene, "Measuring the Segmental Intelligibility of Synthetic Speech: Results from Eight Text-to- Speech Systems," Research on speech perception Progress Report 11, Indiana University, pp. 3-31, 1985.

[27] R.L. Pratt, "Quantifying the Performance of Text-to- Speech Synthesizers," Speech Technology, pp. 54-64, March/April, 1987.

[28] P. Howard-Jones, SOAP, Speech Output Assessment Package, Version 4.0, ESPRIT SAM-UCL-042, 1992.

[29] ITU-T Draft Recommendation P.8S - Subjective Performance Assessment of the Quality of Speech Voice Output Devices. Study Group 12 - Contribution 6, 1993.

[30] U. Jekosch, "The Cluster - Identification Test," ESPRIT SAM Final Report Year Three, II.e, 1992.

[31] U. Jekosch, "Speech Intelligibility Testing: on the Interpretation of Results," Journal of the American Voice I/O Society, Vol. 15, pp. 63-79, 1994.

[32] U. Jekosch and L.C.W. Pols, "A Feature-profile for Application-specific Speech Synthesis Assessment and Evaluation," Proc. 3rd International Conference on Spoken Language Processing, Yokohama, pp. 1319- 1322, 1994

[33] R.D. Johnston, An On-going Series of Subjective Experiments to Assess Epeech Output from Text-to-Speech Systems. Unpublis- hed Report to CCITT Study Group 12, 1993.

[34] W. Jongenburger and R. Van Bezooijen, *Evaluatie van ELK: Attitudes van de Gebruikers, Verstaanbaarheid en Acceptabiliteit van de Spraaksynthese, Bruikbaarheid van het Zoeksysteem.* Stichting Spraaktechnologie, Utrecht, 1992.

[35] W. Jongenburger and V.J. Van Heuven, "Sandhi Processes in Natural and Synthetic Speech," in: *Analysis and Synthesis of Speech, Strategic Research towards High-quality Text-to-Speech Generati- on*, Eds. V.J. Van Heuven and L.C.W. Pols, Mouton de Gruyter, Berlin, pp. 261-276, 1993.

[36] R. Kager, *Stress and Destressing in English and Dutch.* Foris, Dordrecht, 1989.

[37] D.H. Klatt, "The Linguistic Uses of Segmental Duration in English: Acoustic and Perceptual Evidence," J. Acoust. Soc. Am., Vol. 59, pp. 1208-1221, 1976.

[38] S.J. Langeweg, *The Stress System of Dutch.* Doctoral Dissertation, Leiden University, 1988.

[39] J. Laver, J. McAllister, M. McAllister, M. Jack, "A Prolog-Based Automatic Text-to-Phoneme Conversion System for British English," Proc. Second Symposium on Advanced Man-Machine Interface through Spoken Language, November, 19-22, Hawaii, pp. 12-11 ff., 1988.

[40] J. Laver, M. McAllister, J. McAllister, "Pre-processing of Anomalous Text-strings in an Automatic Text-to-Speech System," in: *Studies in the Pronunciation of English: A Commemorative Volume in Memory of A.C. Gimson*, Ed. S. Ramsaran, Croon Helm, London, 1989.

[41] J. Laver, *The Gift of Speech, Papers in the Analysis of Speech and Voice.* Edinburgh University Press, Edinburgh, 1991.

[42] J. Laver, *Principles of Phonetics.* Cambridge University Press, Cambridge, 1994.

[43] P.A. Luce, T.C. Feustel, and D.B. Pisoni, "Capacity Demands in Short-term Memory for Synthetic and Natural Word Lists," Human Factors, Vol. 25, pp. 17-32, 1983.

[44] L.M. Manous, M.J. Dedina, H.C Nusbaum, and D.B. Pisoni, Speeded Sentence Verification of Natural and Synthetic Speech, Research on Speech Perception Progress Report 11, Indiana University, 1985.

[45] A.I.C. Monaghan and D.R. Ladd, "Evaluating Intonation in the CSTR Text-to-Speech System," Proc. ESCA Workshop on Speech I/O Assessment and speech databases, Noordwijkerhout, pp. 3.6.1-3.6.4, 1989.

[46] A.I.C. Monaghan and D.R. Ladd, "Symbolic Output as the Basis for Evaluating Intonation in Text-to-Speech Systems," Speech Communication, Vol. 9, pp. 305-314, 1990.

[47] E. Moulines and F. Charpentier, "Pitch Synchronous Waveform Processing Techniques for Text-to-Speech Synthesis Using Dipho- nes," Speech Communication, Vol. 9, pp. 453-467, 1990.

[48] I.R. Murray and J.L. Arnott, "Toward the Simulation of Emotion in Synthetic Speech: A review of the Literature on Human Vocal Emotion," J. Acoust. Soc. Am., Vol. 93, pp. 1097-1108, 1993.

[49] A.M. Nunn and V.J. Van Heuven, "MORPHON: Lexicon-based Text-to-Phoneme Conversion and Phonological Rules," in: *Analysis and Synthesis of Speech, Strategic Research towards High-quality Text-to-Speech Generation*, Eds. V.J. Van Heuven and L.C.W. Pols, Mouton de Gruyter, Berlin, pp. 88-113, 1993.

[50] H.C. Nusbaum, D.L. Greenspan, and D.B. Pisoni, Perceptual Attention in Monitoring Natural and Synthetic Speech, Research on Speech Perception Progress Report 12, Indiana University, 1986.

[51] P. Nye, J. Hankins, T. Rand, I. Mattingly, and F. Cooper, "A Plan for the Field Evaluation of an Automated Reading System for the Blind," IEEE Transactions on Audio and Electroacoustics, Vol. 21, pp. 265- 268, 1973.

[52] P.W. Nye and J.H. Gaitenby, "The Intelligibility of Synthetic Monosyllabic Words in Short, Syntactically Normal Sentences," Haskins Laboratories Status Report on Speech Research, 37/38, pp. 169-190, 1974.

[53] P.W. Nye, F. Ingemann, and L. Donald, "Synthetic Speech Comprehension: A Comparison of Listener Performances with and Preferences among Different Speech Forms," Haskins Laboratories Status Report on Speech Research, 41, 1975.

[54] M.H. O'Malley and M. Caisse, "How to Evaluate Text- to-Speech Systems," Speech Technology, Vol. 3, pp. 66-75, 1987.

[55] C.V. Pavlovic, M. Rossi, and R. Espesser, "Use of the Magnitude Estimation Technique for Assessing the Performance of Text-to- Speech Synthesis System," J. Acoust. Soc. Am., Vol. 87, pp. 373-381, 1990.

[56] D.B. Pisoni, B.G. Greene, and H.C. Nusbaum, "Some Human Factors Issues in the Perception of Synthetic Speech," Proc. Speech Tech '85, New York, pp. 57-61, 1985.

[57] D.B. Pisoni, H.C. Nusbaum, and B.G. Greene, "Perception of Synthetic Speech Generated by Rule," Proceedings IEEE, Vol. 73, pp. 1665-1676, 1985.

[58] L.C.W. Pols, "Quality Assessment of Text-to-Speech Synthesis-by-Rule," in: *Advances in Speech Signal Processing*, Eds. S. Furui and M.M. Sondhi, Marcel Dekker Inc., New York, pp. 387-416, 1991.

[59] T. Portele, B. Heuft, F. Ho"fer, H. Meyer, and W. Hess, "A New High Quality Speech Synthesis System for German," Proc. Yokohama?/New Paltz? (check with JvS/KKP) ***, 1994

[60] J.V. Ralston, D.B. Pisoni, S.E. Lively, B.G. Greene, and J.W. Mullennix, "Comprehension of Synthetic Speech Produced by Rule: Word Monitoring and Sentence-by-Sentence Listening Times," Human Factors, Vol. 33, pp. 471-491, 1991.

[61] J.C. Roelofs, "Synthetic Speech in Practice: Acceptan- ce and Efficiency," Behaviour and Information Technology, Vol. 6, pp. 403- 410, 1987.

[62] P.J. Scharpff and V.J. Van Heuven, "Effects of Pause Insertion on the Intelligibility of Low

Quality Speech, Proceedings 7th FASE/Speech '88 Symposium, Edinburgh, pp. 261-269, 1988.

[63] E.C. Schwab, H.C. Nusbaum, and B.G. Greene, "Perception of Synthetic Speech Generated by Rule," Proc. IEEE, Vol. 73, pp. 1665-1676, 1985.

[64] C. Sorin, "Towards High-quality Multilingual Text-to- Speech," Proc. CRIM/FORWISS Workshop, Mu"nchen, pp. 53-62, 1994. Also to appear in: *Progress and Prospects in Research and Technology*, Ed. H. Nieman, Infix Publishing Company, Sankt Augustin, 1994.

[65] M.F. Spiegel, M.J. Altom, M.J. Macchi, and K.L. Wallace, "Comprehensive Assessment of the Telephone Intelligibility of Synthesized and Natural Speech," Speech Communication, Vol. 9, pp. 279-291, 1990.

[66] R. Sproat, J. Hirschberg, and D. Yarowsky, "A Corpus- based Synthesizer," Proc. 2nd International Conference on Spoken Language Processing, Banff, Vol. 1, pp. 563-566, 1992.

[67] J.M.B. Terken, (1985) *Use and Function of Accentu- ation.* Some Experiments. Doctoral Dissertation, Leiden University.

[68] J.M.B. Terken, "Human and Synthetic Intonation: A Case Study," in: *Analysis and Synthesis of Speech, Strategic Research towards High-quality Text-to-Speech Generation*, Eds. V.J. Van Heuven and L.C.W. Pols, Mouton de Gruyter, Berlin, pp. 241-259, 1993.

[69] J.M.B. Terken and R. Collier, "Automatic Synthesis of Natural-sounding Intonation for Text-to-Speech Conversion in Dutch," Proc. Eurospeech '89, Vol. 1, pp. 357-359, 1989.

[70] R. Van Bezooijen, "Lay Ratings of Long-term Voice-and-Speech Characteristics," in: *Linguistics in the Netherlands 1986*, Eds. F. Beukema and A. Hulk, Foris, Dordrecht, pp. 1-7, 1986.

[71] R. Van Bezooijen, Evaluation of Two Synthesis Systems for Dutch - Intelligibility and Overall Quality of Initial and Final Consonant Clusters. SPIN-ASSP Report No. 3, Stichting Spraaktechnolo- gie, Utrecht, 1988.

[72] R. Van Bezooijen, "Evaluation of the Suitability of Dutch Text-to-Speech Conversion for Application in a Digital Daily Newspaper," Proc. ESCA Workshop Speech I/O Assessment and Speech Databases. Noordwijkerhout, pp. 6.3.1 - 6.3.4, 1989.

[73] R. Van Bezooijen and W. Jongenburger, "Evaluation of an Electronic Newspaper for the Blind in the Netherlands - Intelligibi- lity, Acceptability, Adequacy, and Users' Attitudes," Proc. ESCA Workshop on Speech and Language Technology for Disabled Persons, Stockholm, pp. 195- 198, 1993.

[74] R. Van Bezooijen and L.C.W. Pols, "Evaluation of a Sentence Accentuation Algorithm for a Dutch Text-to-Speech System," Proc. Eurospeech '89, Paris, Vol. 1, pp. 218-221, 1989.

[75] R. Van Bezooijen and L.C.W. Pols, "Evaluating Text-to-Speech Systems: Some Methodological Aspects," Speech Communication, Vol. 9, pp. 263-270, 1990.

[76] R. Van Bezooijen and L.C.W. Pols, "Evaluation of Text-to-Speech Conversion for Dutch," in: *Analysis and Synthesis of Speech, Strategic Research towards High-quality Text-to-Speech Generation*, Eds. V.J. Van Heuven and L.C.W. Pols, Mouton de Gruyter, Berlin, pp. 339-360, 1993.

[77] R. Van Bezooijen and V.J. Van Heuven, "Assessment of Speech Output Systems: State of the Art and Recommendations," in: *Report of the European Advisory Group on Language Engineering Standards*, Eds. N. Calzolari and J. McNaught, Spoken Language Systems Working Group, chpt. 4 (available through ftp, "nicolet.ilc.pi.cur.it"), 1994.

[78] V.J. Van Heuven and P.J. Scharpff, "Acceptability of Several Speech Pausing Strategies in Low Quality Speech Synthesis; Interac- tion with Intelligibility," Proc. 12th International Congress of Phonetic Sciences, Aix-en-Provence, pp. 458-461, 1991.

[79] Y. Van Holsteijn, "TextScan: A Preprocessing Module for Automatic Text-to-Speech Conversion," in: *Analysis and Synthesis of Speech, Strategic Research towards High-quality Text-to-Speech Generation*, Eds. V.J. Van Heuven and L.C.W. Pols, Mouton de Gruyter, Berlin, pp. 27-41, 1993.

[80] J.P.H. Van Santen, "Perceptual Experiments for Diagnostic Testing of Text-to-Speech Systems," Computer Speech and Langua- ge, Vol. 7, pp. 49-100, 1993.

[81] W.D. Voiers, "Evaluating Processed Speech Using the Diagnostic Rhyme Test," Speech Tech-

nology, Vol. 1, pp. 338-352, 1983.

[82] W.D. Voiers, A.D. Sharpley, and C.J. Hehmsoth, Research on Diagnostic Evaluation of Speech Intelligibility. Research Report AFCRL- 72-0694. Air Force Cambridge Research Laboratories, Bedford, Massachusetts, 1975.

[83] J. Vroomen, R. Collier, S. Mozziconacci, Duration and Intonation in Emotional Speech. Proc. Eurospeech '93, Berlin, Vol. 1, pp. 577-580, 1993.

[84] N. Willems, R. Collier, and J. 't Hart, "Synthesis Scheme for British English Intonation," J. Acoust. Soc. Am., Vol. 84, pp. 1250-1261, 1988.