Chapter 2

EVALUATION OF SPEECH SYNTHESIS

From Reading Machines to Talking Machines

Nick Campbell

National Institute of Information & Communications Technology,
and ATR Spoken Language Communication Labs,
Acoustics and Speech Research Department,
Keihanna Science City, Kyoto, Japan

nick@nict.go.jp

Abstract This chapter discusses the evaluation of speech synthesis. It does not attempt to present an overview of all the techniques that may be used, or to cover the full history of previous evaluations, but instead it highlights some of the weaknesses of previous attempts, and points out areas where future development may be needed. It presents the view that speech synthesis should be judged not as a technology, but as a performance, since the actual intended listener presumably has less interest in the achievements of the developers than in the effect of the speech, its pleasantness, and its suitability to the various listening contexts.

Keywords Speech synthesis; Human communication; Perception; Evaluation; Naturalness; Character; Expression of affect.

1 Introduction

Computer speech synthesis is the art and science of making machines speak or talk. At its inception, the focus of research was on producing reading machines as aids for the handicapped, but nowadays speech synthesis is likely to be encountered in more general situations such as customer care networks, replacing humans in telephone-based services, or providing information such as stock prices, car navigation directions, weather forecasts, and railway-station announcements.

The focus of this chapter is on the evaluation of computer speech systems and components, and it urges consideration of both the aesthetic and the scientific aspects of results when judging the acceptability of computer-generated

L. Dybkjær et al. (eds.), Evaluation of Text and Speech Systems, 29–64.

speech. The technology has now reached a stage at which intelligibility is no longer in question, but how should we now evaluate the naturalness and likeability of synthesised speech? It is useful to consider here how we judge human speech; certainly not from the point of view of its technological achievements – those we take for granted – but rather by its effects. Perhaps by its expressivity?

We should start by defining some evaluation criteria. Here too, human speech might offer some useful parallels. Parents judge the developing speech of an infant by its intelligibility, recognising words amongst the babble, but computer speech comes already endowed with a full set of lexical rules from the start. Teachers judge the quality of foreign-learners' speech from a combination of intelligibility and naturalness, considering prosodic appropriacy alongside linguistic intelligibility. Actors judge prosodic expression first of all, determining whether the speech portrays the intended deeper meanings of its text when performed, and newscasters judge the personality as well as the expressiveness in a speaker's voice, evaluating not just the fit between words and expression but also the character of the speaker and whether the "tone" is appropriate for the desired station image. These are harsh criteria for evaluating computer speech, but unless we start from a framework in which human speech is likely to be judged, we are in danger of accepting unsatisfactory aspects of speech production and endorsing products that will be likely to cause frustration or dissatisfaction in the human listener.

2 Components of Computer Speech

It is generally agreed that there are three main stages in the production of computer-generated speech, whether from text or from concept. The first is the language-processing stage; producing a machine-readable representation of the input in a form that indicates both (a) the word sequence and its pronunciations, and (b) the relations between the words so that their intended meaning can be understood. The second stage is prosody processing, converting the abstract text-based representation of the speech into a sequence of parameters representing the pitch, energy, duration, and voice quality of each acoustic segment. The third stage is waveform generation, a component that takes the parameter-based representation and converts it into a sequence of waveform segments that can be sent to an audio device for presentation to the listener.

It is common to evaluate each component stage separately, but it is also necessary to perform a holistic evaluation, since the interactions and dependencies between each component can also have an important effect on the acceptability of the resulting speech.

2.1 Text Pre-Processing

Producing a machine-readable representation of the input in a form that indicates both the word sequence in pronouncable form and the relations between the words.

The input to a computer speech synthesiser is commonly in the form of plain text. This text may be pre-existing in the outside world in the form of newspaper articles, books, etc., or from keyboard input, or it may be machine-generated, as in concept-to-speech systems.

In the former case, of pre-existing text, the speech-related information must be generated by the synthesis system. This processing takes two main forms: anomaly resolution and dependency relations.

In the latter case, the text does not have the requirement of being human-readable, and can be pre-annotated with various forms of speech-related information to aid in its disambiguation, pronunciation, and phrasing, in much the same way that the layout and format of a web page can be specified by the use of XML markup and style sheets.

The Speech Synthesis Markup Language (SSML) Version 1.0 home page (http://www.w3.org/TR/speech-synthesis/) of the World Wide Web Consortium summarises this goal as follows:

> The Voice Browser Working Group has sought to develop standards to enable access to the Web using spoken interaction. The Speech Synthesis Markup Language Specification is one of these standards and is designed to provide a rich, XML-based markup language for assisting the generation of synthetic speech in Web and other applications. The essential role of the markup language is to provide authors of synthesizable content a standard way to control aspects of speech such as pronunciation, volume, pitch, rate, etc. across different synthesis-capable platforms.

Note that no effort has yet been made to incorporate paralinguistic information, except by direct manipulation of the lowest-level acoustic parameters, making specification by the non-specialist rather difficult.

2.1.1 Making sense of the text: anomaly resolution.

Not all text that is clear to the eye is so easily converted into sounds; abbreviations such as "Mr." and "Dr." may be ambiguous (the latter, for example, typically representing "doctor" when coming before a proper name, and "drive" when coming after one) and common acronyms may require different pronunciation rules (IBM being pronounced as three separate letters, and JAL being pronounced as a single monosyllabic word, for example). Such textual anomalies must be converted into phonetic representations in this component of the speech synthesiser. Dictionary entries listing the possible pronunciations require complex rules to determine which is most appropriate in any given case.

Dependencies between words can require even more complex knowledge before their proper pronunciation can be determined; the simple text sequence "old men and women", for example, has a potential ambiguity in the scope of the adjective "old": does it refer to the men alone, or to both the men and the women? The former interpretation would require a short pause after "men", and a pitch reset before "women", but the latter would require that the

words be reproduced as a single phrase with no prosodic reset. Considerable text analysis (and some world knowledge, usually represented in the form of statistical probabilities) is required before an appropriate pronunciation can be determined.

2.1.2 Differences between text and speech. Text and speech have evolved differently and independently throughout human history, but there is nonetheless a common misconception that text can be simply converted into speech (and vice versa) through a process of media transformation, and that the two media differ primarily only in terms of their surface form. Speech, however, is by far the older medium (in terms of human evolution) and is usually conversational and interactive, one-to-one, directly signalling many more types of information than text does.

As well as carrying propositional content (and often instead of), speech contains many clues about the speaker's affective states, intentions, emotions, identity, health, and both short- and long-term relationships with the listener. Speech incorporates paralinguistic and extra-linguistic information that is not present (or relevant) in text. Text, on the other hand, has evolved for the eye rather than for the ear; being two-dimensional, it can be scanned from top to bottom, and from right to left, remaining present and (usually) unchanging on the page or screen to allow a more considered and analytical processing of its content. We typically scan more than one word at a time when reading text, but we receive spoken input linearly, as a one-dimensional sequence of sounds.

The speech signal decays with time, but speech makes full use of prosody to carry the information of this missing second dimension. The complex syntactic and semantic relations of a text must be converted into prosodic information before rendering it into speech, but this transformation is something which even many humans find difficult. The art of reading a text aloud requires considerable training, and very few "ordinary people" can do it well. Yet we expect a perfect enunciation from a speech synthesiser! Perhaps text-to-speech pre-processing remains the most challenging component of speech synthesis, but it can be done well only in the most uninteresting of cases. However, with annotated input, or marked-up text, the responsibility for determining the surface realisation of the utterance is passed back to the human author of the text, who has the best understanding of what the desired output rendering of each utterance should be.

2.2 Prosody Prediction

Converting the abstract representation of the speech into a sequence of parameters representing the pitch, energy, duration, and voice quality of each segment.

Prosody in speech is used to portray a complex bundle of information, the strands of which also include speaker identity, character, health, mood, interest, and emotion. At the same time, it serves also to distinguish nouns from verbs (e.g., "import", "record"), proper nouns from common nouns (e.g., "White House" vs "white house"), and to show syntactic relations and phrasing by means of pauses and pitch resets. The calculation of speech prosody for a given utterance is therefore a major component of speech synthesis, but one which relies on receiving adequate input from the text-processing component that precedes it.

2.2.1 Knowing the right answer. One of the difficulties in evaluating the prosody of computer speech synthesis is that it requires prior knowledge of the intended interpretation of any synthesised utterance. Many prosodic contours may be appropriate, but only one of them correct for a given utterance in context. However, since sufficient world knowledge or context-related information is not always available to the text- and prosody-processing components, a so-called default prosody is often generated instead. This is the least-marked form that an utterance can take. The use of a so-called default may render mistaken interpretations less likely, but it also results in the flat and monotonous tone of much synthesised speech.

2.2.2 Four prosodic variables. Prosody is typically realised first through manipulation of the fundamental frequency contour of a synthesised utterance. The second major prosodic variable is segmental duration, which can function instead of, or in conjunction with, the fundamental frequency contour to signal the intended interpretations of an utterance.

The manipulation of either pitch or duration requires either signal processing of the speech waveform (which can have damaging effects on its naturalness) or judicious selection of segments for concatenation (which in turn requires an inordinately large repository of sounds from which to make the selection).

Little research attention has been given to prosodic manipulation of the energy contour (traditionally considered to be the third prosodic variable), partly because it correlates quite closely with the fundamental-frequency information, and partly because the voice level of synthesised speech is usually held at a constant predetermined volume.

The fourth prosodic component, which is recently receiving more attention in the speech research communities, is voice quality, or manner of articulatory phonation, which can be varied to indicate emotion or speaker activation and interest, in addition to marking focus or emphasis and to showing characteristics related to phrasal and utterance position.

2.2.3 Acceptable variation in prosody. Whereas the output of a text-pre-processing module can perhaps be evaluated in isolation, without resorting to synthesis, it is more difficult to evaluate prosody prediction in this way. There can be large differences in the physical measurement of the prosodic parameters that do not appear as perceptually relevant differences in the realisation of an utterance as speech. An obvious example is the difference in pitch (or more correctly fundamental frequency range) between the speech of men, women, and children, which human listeners appear to disregard unconsciously, since they pay less attention to these individual differences and hear instead the "intended prosody" of an utterance regardless of often very large differences in acoustic prosodic parameter values. This is not a claim that people do not hear these large differences, but rather that they can perceive utterances that have the same content as "the same" even though they may be spoken by different people.

A less obvious example is the raising or lowering of pitch in the speech of a single speaker; if all values for a given utterance are raised or lowered by the same amount, and if that amount is still within the normal speaking range of that speaker, then no difference in intended meaning will be perceived. Conversely, even a slight delay or advancement of a pitch peak in an utterance can result in a profound difference in perceived meaning of the utterance. Thus, simple numerical measures of, for example, difference between the prosody of a natural utterance and that of a synthesised one can be extremely difficult for the evaluator to interpret.

Furthermore, we observe considerable differences in the absolute values of the prosodic parameters in the speech of humans, all of which would be considered "natural" and "correct" by a listener, but often a much smaller difference in the output of a speech synthesiser is perceived as quite unnatural, or as presenting a different interpretation. This paradox is not yet well explained scientifically.

2.3 Waveform Generation

Taking the parameter-based representation and converting it into a sequence of waveform segments that can be sent to an audio device for presentation to the listener.

There are many ways to produce speech waveforms by computer; the simplest (perhaps) is by concatenation, just joining together segments selected from different parts of a larger database, and the most complex is by modelling the variations in the articulatory tract and voice-source parameters and then reproducing them by rule. The science of acoustics and the physics of wave propagation find practical application in this area, but psychology has yet to enter.

2.3.1 From phones to segments. The earliest attempts at computer modelling of speech employed formant approximations of the phones (and produced a smoothed interpolation between them) to replicate the acoustics of vowel and consonant differences as observed from spectrographic representations of human speech utterances. The works of Gunnar Fant, Ken Stevens, and Dennis Klatt contributed considerably to our understanding of the physics of speech production and formed the basis of much of modern speech synthesis.

Later contributions, most notably from Osamu Fujimura, Joe Olive, and Yoshinori Sagisaka, resulted in a different, non-phonemic view of speech segmentation, and led to the concatenative systems that are in more popular use today. Rather than modelling the "centres" of speech information and interpolating between them, they recorded the "transitions" (i.e., made use of dynamic rather than static information) and concatenated short segments of speech waveforms to produce the synthesised speech.

More recent improvements in concatenative speech synthesis have resulted in less need (often none at all) for signal processing to modify the recorded waveform segments, by enlarging the source-unit database, enriching it, instead of using a smaller unit-inventory and resorting to signal modifications to manipulate the prosody.

2.3.2 System size and output quality. Whereas articulatory modelling of speech may offer the most scientific benefit, it is still very far from being of practical use in real-time speech synthesis systems, and concatenative synthesis appears to offer a higher quality of speech output for the present time, leaving parametric synthesis (such as formant-based synthesis-by-rule) to offer a lower-quality but smaller-footprint synthesiser that may be more suitable for use in applications where speech quality is of lesser importance than price or memory requirements.

Of course, each of these synthesis systems has its own individual strengths and weaknesses, and each should be evaluated in situ, taking into account not just the overall quality of the resulting speech, but also its appropriateness for a given synthesis application. We should not think in terms of "one best system", but of a range of different methods that can be selected from, and possibly even switched between, according to the needs and preferences of the individual user.

Paradoxically, the smallest system may offer the most flexibility, being able to mimic many voices and speaking styles but at the cost of not being able to sound "human", and the most expensive (in terms of speech segment inventories) may be limited to the voice of only one speaker and one speaking style, though able to reproduce that with such a precision that it may be difficult or sometimes impossible to notice any difference from natural human speech.

Even among human speakers, there is a similar trade-off of talents, as very few professional speakers are able to master all modes of presentation equally well. A Shakespearean actor might make a very poor newsreader, for example, yet it is often assumed that computer speech synthesis will one day outperform them all! This is still an unrealistic assumption, and we should be looking instead for different methods that can match different expectations, rather than one global system that can excel at all.

3 Evaluation Methodologies

There are many different ways that a speech synthesiser can be evaluated: diagnostic or comparative, subjective or objective, modular or global, task-based or generic, etc., and probably as many different ways to perform the evaluation: web-based or live, over headphones or loudspeakers, with actual users or recruited listeners, specialist or naive, pairwise or in isolation, and whether or not human speech samples are included in the same evaluation with the synthesized speech.

Early speech synthesisers were evaluated primarily for intelligibility, using rhyme tests, anomalous sentences (e.g., the Haskins set which had the form "the *adj noun verb* the *noun*") and lists of words both in sentences and in isolation. The goal of such evaluation was phonetic discrimination, i.e., focussing on "intelligibility" rather than "naturalness".

Intelligibility is of course necessary, but naturalness is perhaps also desirable. As early as 1974, this point was being addressed:

> From our point of view it is not physical realism but psychological acceptability which is the proper evidence for correctness at the phonological and phonetic levels, just as it is on the syntactic level. *(I.G. Mattingly, 1974. Developing models of human speech.)*

However, whereas "naturalness" may be difficult to judge in synthesised speech, different synthesisers (or different settings of a given synthesiser) can be compared instead for *relative* naturalness. The need for human realism is variable; in some applications it is essential, while in others it is undesirable. Yet in all cases, the listener must endure the speech, so I here propose a third evaluation criterion: "likeability", in addition to the standard two above, as a measure of the degree to which extended listening becomes bearable, or even perhaps enjoyable.

3.1 Component-Based Evaluation

Whereas the listener (or "customer" perhaps) is primarily concerned with the overall quality, intelligibility, and likeability of the output speech, the developer is usually more concerned with testing one component at a time. Not all components can be run in isolation, so contrived input may be necessary for testing purposes, but in judging a speech synthesis system as a whole, it can be

difficult to determine the source of any specific error, and small errors may be compounded as they cascade through later components.

3.1.1 Evaluating the text pre-processing component.

Because the mapping between a text and its pronunciation is deterministic, this component of the speech synthesiser is, paradoxically, perhaps the easiest to evaluate using objective measures. For example, the part of speech of a given word in a given context is a fact that may be difficult to estimate from limited knowledge in a computer program, but one which has a single right answer that can be checked.

Similarly, the pronunciation of a text sequence might vary with, for example, speaking rate, or dialect, but it is in general predetermined. For example, the letter sequence /b-a-s-s/ will be pronounced one way in the context of music, and another in the context of fishing, but in either case there is a "right" pronunciation. This is not the case with proper names, however, which can vary even according to personal whim, and for which no dictionary or rule set can provide a definitive answer.

Number strings require context-specific rules for their realisation as word sequences: telephone numbers having a different syntax from ordinals, as do special symbols (e.g., $N = "dollar-N" in a computer program, but "N-dollars" in a financial context), but again, these can be categorically judged for correctness.

Sometimes the input to the text preprocessor is by default ill-formed. The listings in a telephone directory, for example, are usually highly abbreviated and, although intelligible to a human reader, may be very difficult to disambiguate by rule, but again, there is usually one "right answer".

Dictionaries can be produced, and letter-to-sound rules trained on them so that only the exceptions need to be stored. The output of these rules, fall-back dictionaries, parsers, and morphological analysers can be evaluated objectively for a large subset of the language (that which excludes proper names, for example) and scores can be calculated to evaluate the component performance.

However, the remaining (almost infinite) subset of texts which contain lexical items whose pronunciation may be arbitrary, can only be evaluated subjectively. Since even humans will not always agree on, for example, the pronunciation of proper names, there can be no objective measure and scoring for "acceptability" rather than correctness will be needed. This work then becomes labour-intensive and expensive.

If the results are passed through the remaining modules of a synthesiser so that the test can be performed aurally, perhaps even over a telephone line, then the output is likely to be corrupted further by compounding of errors as they cascade through the system. However tedious it may be to perform, a text-based examination of the component output is therefore preferable to listening tests for evaluating the early stage of speech synthesis processing.

3.1.2 Evaluating the prosody component. Given perfect input, the prosody component should be able to reliably produce the specifications for an unambiguous rendering of the text as speech when output by the waveform component. Its input can be modified if necessary, to represent the perfect output of a text pre-processing stage, but how is its output to be evaluated?

As mentioned above, there is tremendous variability in the prosody of human speech, much of which is perceptually insignificant or cognitively irrelevant, but some of which has the effect of changing the entire meaning of an utterance. The word "yes" for example, if spoken slowly and with a rise-fall-rise contour, can even signal "no" to a familiar listener.

Textual anomalies, such as "I saw Janet and John saw Fred" (where "Janet and John" might be a compound subject) must be resolved at an earlier stage and the disambiguating information passed as input, but in order to judge the correctness of predictions from the prosody component, the listener (or reviewer) must first know the *intended interpretation* of the utterance.

It is highly unlikely that any two speakers will produce the "same" utterance with identical prosodic contours, yet most human listeners would be able to judge whether or not there was an intended difference in meaning or nuance between any pair of utterances from different speakers. As a science, we do not yet have enough understanding of this flexible variability in perception or production to produce a model or technology that would enable its objective evaluation. Yet to assess the performance of a prosody component using aural perception requires passing it through a further stage of processing in which errors can be compounded (or hidden). This is the "catch-22" of prosody research; we need to know about both the speaker's intentions and the listener's perceptions of them, yet only in a very few "black-and-white" cases can these be categorically determined.

When we expand our goals beyond judging correctness, towards judging paralinguistic expression, the problem becomes even more difficult, but if a computer speech synthesiser is to be used in place of a human speaker in interactive discourse situations, the expression of paralinguistic and extra-linguistic information is as important (if not more so) as the expression of linguistic or propositional content.

Teams of listeners are required for an efficient evaluation, and in the case of the commonly used Mean Opinion Score (MOS) evaluations, a minimum of 30 listeners is necessary for a statistically significant result. ABX tests are common (comparing two versions to a target and indicating the closer one) as are preference tests, but it is very difficult for a listener to precisely identify the particular cause of a prosodic problem as the four elements of prosody are perceived as an integrated whole, and events occur too quickly for the ear to be able to identify their exact location. Diagnostic testing of prosody is therefore a difficult art.

3.1.3 Evaluating the waveform component. In earlier times, the focus of output evaluation was on mechanical performance and particularly on the ability to mimic human phonetic sequences, so rhyme tests and dictation-type exercises were adequate to determine if a target phonetic sequence was correctly perceived. Nowadays, though, as technology progresses, we advance beyond segmental intelligibility and are more concerned with judging expressiveness and personality in computer speech.

The quality of audio on the web has now reached hi-fi performance but the quality of most synthetic speech is already lagging quite far behind. Again, the needs with respect to quality must be regarded as application-specific, and it is probably not necessary to provide broadcast-quality speech in a talking wristwatch or toaster, but the failure to meet the expectations of the ordinary person-in-the-street might be the main reason that speech synthesis technology has been so slow in getting accepted.

When evaluating the waveform generation component, we can assume that the rest of the system has functioned perfectly and provide it with (perhaps hand-crafted) ideal input as well as with input that has actually been generated by passing through the earlier components. In an ideal world, there will be no difference between the two, but any difference that is found at this stage can be attributed to earlier processing inadequacies and ignored for the sake of the evaluation.

Physical measures of waveform similarity are used in speech recognition and can be used similarly in speech synthesis for determining an objective measure of the distance between one speech waveform and another. Unfortunately, none of these measures matches human perception perfectly, and they either underestimate problems that human listeners might find noticeable, or raise an error on points that the human listener might not notice. If a measure that perfectly matched human perception could be found, the problems of concatenative speech synthesis would be over, for the minimisation of error could then be performed automatically.

As with prosody evaluation, waveform generation is best evaluated by subjective measures, and again, MOS (which is widely used in the telecommunications industry for speech-coder assessment) has proved to be a very good indicator of overall performance.

3.2 Evaluating the Complete System

Component evaluation is a necessary part of system development, and provides both feedback and diagnostic information to the researcher, but even if every component is working well, their integration can sometimes lead to problems, so a full system-level evaluation is also required. Of prime importance is intelligibility of the output, then its naturalness, and last, but by no means least, its likeability. People may buy synthesisers "off the shelf", and first impressions will have a strong influence on their decisions, but after repeated use, the

character of the system becomes as important as its clarity. If people do not like a synthesiser's voice and speaking styles, they will soon cease to use it.

3.2.1 Intelligibility. Tests of nonsense words and semantically anomalous sentences do provide a measure of the intelligibility of synthesised speech output, but not one that realistically represents everyday situated use. If the tests were carried out in a car, a helicopter cockpit, a busy shopping centre, or at home with a television and children playing in the background, they might be more useful, but this is rarely the case. Researchers seem to be more interested in evaluating topline performance than baseline reliability.

Now that the science of speech synthesis has progressed beyond the early stage of testing phonetic adequacy, perhaps we will begin to see more task-based evaluation. There have been some performance-based measures of fatigue and response time delays related to use of speech synthesisers in the performance of a real-world task, but this mode of testing, although it better reflects the actual needs of the end-user, seems to be the exception rather than the rule.

As speech synthesis becomes more expressive, particularly when used in place of a human voice, as in translation systems and communication aids, the variability of speaking style will require fast and even slurred speech to be produced if it is to faithfully represent the intentions of the original speaker. It is not yet known whether the present waveform generation methods will be robust against such intentional distortion of the clean-speech signal, so quantitative measures of degradation in intelligibility (both segmental and prosodic) will also be required.

3.2.2 Naturalness or believability. Perhaps "believability" would be a better term to use instead of "naturalness", although it has not been seen much in the previous literature, because even an artificial-sounding voice (as in fictional cartoons, for example) can sound believably natural if the prosody is appropriate.

The world of Walt Disney is a perfect example of why "naturalness" may not be the best way of approaching this aspect of speech synthesis evaluation. Very few of the creatures in Disney films are photo-realistic, yet they are all very believable. So how should we measure believability in speech synthesis?

Fitting the voice and speaking style to the content and context of the message requires a flexibility in all three stages of synthesis design. The success of such an effort can be intuitively felt by many listeners, even if they cannot explicitly quantify it. The selection of telephone personnel, newsreaders, actors for roles, and characters in advertisements, is a very sensitive process involving delicate judgements about voice as well as role suitability and matching. This framework should be adapted for use in synthesis evaluation.

3.2.3 Likeability. The selection of an ideal voice for use in a concatenative speech synthesiser is a very lengthy process. Not only must the voice have characteristics that allow for easy joining between segments extracted from different parts of the database, but it must also have an appeal to the listener and a fit to the perceived or projected image of the company or product whose voice it is to become.

Whether produced by concatenation or signal processing, whether from a large source database or a small one, the resulting voice and the range of speaking styles that it is capable of will define the popularity of a synthesiser as much as technical ability in the processing of the various stages.

Current evaluation methodologies are biased towards accountability and reproducibility, favouring objective measures over subjective ones, but ultimately, the listener's feelings towards the synthesiser must also be allowed to play a role in the evaluation process. Likeability though is an aspect of synthesiser use that can easily change over time. What sounds fun and attractive on first listening may become irritating after prolonged use, or what is appealing in a quiet environment may become inadequate when used in noise. Long-term assessment is an issue that must be addressed if we are to produce a technology that is pleasant as well as efficient.

4 Organised Evaluations and Assessment

In this section we will examine some techniques that have been successfully used in the past and suggest some new ones that may better address the developing needs of the current technology.

4.1 Learning from the Past

Many scientific presentations concerning speech synthesis are accompanied by demonstrations of the resulting output, but these are usually demonstrations of carefully controlled samples that are limited in number and that do not necessarily give a fair impression of the overall performance of a given synthesis system as it might perform across a wide range of everyday application tasks.

Scientific research is rightly more concerned with raising the topline of achievements, as evidenced by the fact that negative findings are rarely reported at academic meetings, but the products of developers are evaluated by the customers throughout their full range of everyday use. While it is topline performance that is demonstrated at conferences, the owner/user of a synthesis system is usually more concerned about the baseline; i.e., not just how well a system can perform in the best of cases, but also how badly it might perform in the worst. A large part of the role of evaluation is to bridge this gap between scientists and developers and to bring the thoughts of the end-user into an earlier stage of the design process.

Evaluation of speech synthesis is by no means a well developed art. In comparison with the number of papers published reporting advances in speech synthesis, the number of papers describing how these advances might be evaluated is still remarkably few. At the recent ISCA Speech Synthesis Tutorial and Research Workshops, for example, there has been on average only 4% of papers explicitly addressing evaluation issues (SSW5, 2004: 3/46; SSW4, 2001: 1/62; SSW3, 1998: 3/60).

4.1.1 The Jenolan experience. At the ESCA/COCOSDA Tutorial and Research Workshop on Speech Synthesis (SSW3), held in 1998 at Jenolan in the Blue Mountains near Sydney, Australia, a concerted evaluation of the then current speech synthesis systems was organised in an attempt to remedy this imbalance. The evaluation served three goals:

1. To give the participants first-hand experience with a wide range of current text-to-speech systems (altogether 70 system/language combinations were presented)
2. To stimulate discussion of speech synthesis evaluation procedures by offering direct experience to contributing participants, who function both as developers and as evaluators
3. To provide feedback to the system developers about some of the strong and weak points of their speech synthesisers

The evaluation was limited to full text-to-speech systems but employed several text types, ranging from newspaper text through semantically unpredictable sentences to telephone directory listings. The newspaper text sentences were further subdivided into "easy', using frequent words, and "hard", using sentence selection based on trigram-frequencies calculated from the Linguistic Data Consortium (LDC) text corpora.

The texts were automatically generated from large corpora for several languages (though not all), and were made available to participants via an interactive web site a very short time before the evaluation was carried out. Speech files had to be synthesised from the texts and returned via the same site to the organisers who then randomised them for presentation to participants at the workshop. Regardless of original sampling rate, all submitted waveform files were resampled to 11.25 kHz by the evaluation organisers before the comparisons were performed.

All participants in the workshop took part as participants in the listening tests. Utterances were presented to the listeners via headphones, and responses were entered using rating scales (from "poor" to "excellent") and also identifying problem areas. Since the listeners were mainly professional synthesis researchers, advantage was taken of their expertise for free input in the latter category, while the former included such scales as "overall voice quality",

"naturalness", "wrong syllable stressed". Experimental design was blocked by text type within listeners, all listening to each text-to-speech system exactly once with each text item for a given language.

Full details of the evaluation procedure are still available at the original COCOSDA web site (http://www.slt.atr.co.jp/cocosda/evaltext.htm), but the results of the evaluation remain confidential as this became a condition of participation. There was concern that (a) system developers would refrain from participating otherwise, and (b) that the results of such a (necessarily) rushed evaluation might not be completely reliable. However, although the experience gained both by organisers and participants was considered valuable, the experiment has not been repeated.

4.2 Synthesis Assessment and Evaluation Centres

There is now no shortage of suitable organisations that could take on the role of assessment centres for synthesis systems in the various languages and world regions, but there is at present no such centre and currently no such effort that the author is aware of.

The LDC was very helpful in providing texts for the Jenolan evaluation and for a while afterwards the LDC maintained a speech synthesis evaluation web site to enable similar web-based comparisons. Unfortunately, at the present time, this site (the interactive speech synthesizer comparison site at http://www.ldc.upenn.edu/ltts/) does not appear to be supported.

The European Language Resources Association (ELRA; http://www.elda.fr/) is a sister organisation to the LDC, based in Europe. Their distribution agency (ELDA) actively participates in evaluation projects, and is beginning to include speech synthesis:

> In the near future, as we are getting more and more involved in the evaluation activity, ELRA & ELDA will add in its catalogue further resources and tools related to evaluation, and a new team, whose task will include the organisation of evaluation campaigns and every other aspects of the evaluation activity, will join the agency. (2005)

The International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques for Speech Input/Output (COCOSDA; http://www.cocosda.org) organised the Jenolan Assessment, in conjunction with the International Speech Communication Association (ISCA (then ESCA); http://www.isca-speech.org/), but after more than 10 years of activity, this group, too, no longer participates in evaluation programmes, though their annual workshops stress the importance of assessment as part of technology development and form a central meeting place for such discussions and information exchange.

A spin-off from COCOSDA, the ISCA Special Interest Group for Speech Synthesis (SynSIG; http://feast.his.atr.jp/synsig/) has taken over the task of organising tutorial workshops related to speech synthesis, and would be the most likely independent (and fully international) organisation to coordinate future speech synthesis assessments, but a repeat of Jenolan is unlikely unless a committed team of volunteers is formed to ensure that the meticulous organisation and arrangement of details can be repeated.

In the Far East, the ChineseLDC (http://www.chineseldc.org/) and the GSK in Japan (Gengo Shigen Kyoyukiko is Japanese for Language Resource Consortium) both aim at developing, sharing, and distributing language resources for research into speech and natural language processing. Both groups are committed to facilitating speech research and will perhaps also take on an assessment and evaluation role in the near future.

4.3 More Recent Assessments

The Verbmobil project across Germany and the international TC-Star project for speech-to-speech translation both conducted end-to-end evaluations that included a speech synthesis component. Unfortunately, most speech-to-speech translation systems have yet to take expressiveness of the input speech as a relevant factor in the translation, and in both cases the output speech synthesis was considered to be a relatively minor component, with the bulk of the evaluations being performed at the level of text. Although both systems accepted human speech as input, as long as the translated text was rendered intelligibly the synthesis was considered to be satisfactory, and little attention was paid to voice quality or expressiveness of the output per se.

The European project COST 258 (http://www.icp.inpg.fr/cost258/), lasting from 1997 to 2001, was concerned with the naturalness of synthetic speech in concrete applications, with a particular focus on the improvements of sound quality and prosodic modelling. It recognised four priority topics for speech synthesis evaluation:

1. Prosodic and acoustic effects of focus and/or emphasis
2. Prosodic effects of speaking styles
3. Rhythm: what is rhythm, and how can it be synthesised?
4. Mark-up: what prosodic markers are needed at a linguistic (phonological) level?

Their book (published by Wiley, see below) can be considered essential reading on the topic.

Recently, from the United States and Japan, the Blizzard Challenge (http://www.festvox.org/blizzard) for evaluating corpus-based speech synthesis using

common databases is becoming an example of the direction that future bottom-up assessment initiatives might take. The goal of this challenge is for different groups to each use the same publicly available speech databases to build a synthetic voices. Unknown sentences from an independent source will be generated and each participant will synthesize them with their system. The synthesised speech will then be put on the web for evaluation. The results are not available at this time of writing, but such an open evaluation of different methodologies using a common resource in the public arena is bound to be helpful.

5 Speaking to (and on Behalf of) People

To date, the knowledge underlying speech synthesis research has been that of phonetic science and its acoustic correlates, and there has perhaps been an undue emphasis upon the linguistic and segmental components of spoken language. Speech production per se rather than *communication* has been the goal. Yet in many of the current applications of speech synthesis, we find that machines are now acting in the place of people to impart information in an interactive conversational framework.

Recent studies of conversational speech have shown that only a small part of the total speech activity is devoted to the pure expression of propositional content, and that the larger part is devoted to maintaining successful discourse flow, expression of affect, expression of speaker–listener relationships, and revealing the speaker's attitude towards the content of the message. These expressions of paralinguistic information require more than just linguistically well-formed utterances, and place stricter demands for more sophisticated uses of prosody than current speech synthesis systems are capable of.

Synthesisers have traditionally been regarded as reading machines, and the term "text-to-speech" is often thought of as synonymous with "speech synthesis", but the emphasis is changing, and "talking machines" are beginning to appear. These machines will not just be required to talk, but also to laugh (for sure) and even to cry perhaps. They will be required to express more than just emotion, and to take part in an interaction where not just the content of the message is of importance, but also the tone of voice, and the manner of speaking. Computer speech will have to brush up on its interpersonal skills if it is to keep pace with the changes in society and technology.

6 Conclusion

This chapter has presented an overview of some of the main issues concerning current speech synthesis assessment and evaluation. Although there has been a tremendous amount of research into this subject in the past, it seems

that the goalposts are always moving and that as the technology advances, so must the evaluation methodologies. Unfortunately there may still be a mismatch between the goals of the scientists and technicians who are developing the systems and the needs of the people-in-the-street who we hope will buy and benefit from their products. In spite of more than half a century of advanced research, few ordinary people yet use a speech synthesiser in their daily lives; at least, not unless they have to. This is the ultimate assessment of our technology.

Part of the problem is that we, as developers, necessarily see the system in terms of its component parts and then evaluate the performance of each on a limited set of criteria that are defined in terms of technological and methodological goals, rather than holistic perception targets. Speech is broken down into processes and sub-processes, and yet in its entirety, speech is more than the just sum of these parts.

What is missing from current speech synthesis technology is a model of interpersonal communication strategies. We have reached the stage at which we can successfully mimic and control many of the individual speech production processes, but we lack a way of evaluating how well they can be used to signal all of the integrated factors that combine to form human speech. We long ago mastered control of the linguistic component, and recently succeeded in incorporating extra-linguistic, speaker-specific information into the technology. What is still missing is a control of paralinguistic information, and a means to evaluate the subtle nuances and shifts of meaning that add richness to human speech. The challenge of modelling global weather systems and climatic effects is known to be a very difficult task, yet understanding and modelling the subtlety and beauty of human speech processes is a far more challenging one. It will perhaps keep us busy for many years to come.

Acknowledgements The author wishes to express thanks to the National Institute of Information & Communications Technology, Japan, to the Japan Science & Technology Agency, and to the management and staff of the Advanced Telecommunications Research Institute in Kyoto for their kind support of this work.

Further Reading

There are already several excellent books on speech synthesis that the interested reader could use as a source for further information, though there is none that I know of which is dedicated solely to the topic of evaluation and assessment.

A comprehensive manual that does include much relevant information is the *Handbook of Standards and Resources for Spoken Language Systems*, by Dafydd Gibbon, Roger Moore, and Richard Winski (editors), published by Mouton de Gruyter (November 1997).

A collection of extended papers from the first ISCA ETRW on Speech Synthesis can be found in *Talking Machines: Theories, Models, and Designs*, by G. Bailly, C. Benoit, and T.R. Sawallis, published by North-Holland (1 May 1992) and a follow-up *Progress in Speech Synthesis*, by Jan P.H. Van Santen, Richard W. Sproat, Joseph P. Olive, and Julia Hirschberg, published by Springer-Verlag (15 January 1997).

No books have resulted from the subsequent ITRW speech synthesis workshops, but their proceedings are available on the web under the home pages of ISCA (http://isca-speech.org). For a more general introduction to the field, the following are recommended:

An Introduction to Text-to-Speech Synthesis, by Thierry Dutoit (Faculté Polytechnique de Mons), published by Kluwer Academic Publishers (Text, Speech and Language Technology series, edited by Nancy Ide and Jean Véronis, volume 3) (1997). *Improvements in Speech Synthesis: Cost 258: The Naturalness of Synthetic Speech,* by E. Keller, G. Bailly, A. Monaghan, J. Terken, and M. Huckvale (editors), published by Wiley (November 2001). *Text to Speech Synthesis: New Paradigms and Advances,* by Shrikanth Narayanan and Abeer Alwan, published by Prentice Hall IMSC Press Multimedia Series (2004).

For the specialist reader, the proceedings of the following workshops and conferences will contain papers of interest:

LREC, the International Language Resources and Evaluation Conference

ICASSP, the International Conference on Acoustics, Speech, and Signal Processing

Eurospeech, the European Conference on Speech Communication and Technology

ICSLP, the International Conference on Spoken Language Processing

ISCA ETRWs, the series of Speech Synthesis workshops organised by ISCA

COCOSDA, International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques for Speech Input/Output.

Although a Google search will probably be more useful to the reader since it can be interactively directed and presents the very latest information, the following reference section summarises previous work that may not appear on the web. It presents a small sample of the types of research that have been

conducted on the evaluation of speech synthesis systems, and in spite of this tremendous body of knowledge, we can still say that the science of speech synthesis evaluation is just at its beginning, and that the effective evaluation of synthesis systems has probably not even begun.

References

Akers, G. and Lennig, M. (1985). Intonation in Text-to-Speech Synthesis: Evaluation of Algorithms. *Journal of the Acoustical Society of America*, 77(6):2157–2165.

Benoit, C. and Pols, L. C. W. (1992). On the Assessment of Synthetic Speech. In Bailly, G., Benoit, C., and Sawallis, T., editors, *Talking Machines: Theories, Models and Designs*, pages 435–441, Elsevier, North Holland, Amsterdam, The Netherlands.

Benoit, C., van Erp, A., Grice, M., Hazan, V., and Jekosch, U. (1989). Multilingual Synthesiser Assessment Using Semantically Unpredictable Sentences. In *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech)*, volume 2, pages 633–636, Paris, France.

Bernstein, J. (1982). Evaluating Synthetic Speech. In *Proceedings of the NBS Workshop on Standardization for Speech I/O Technology*, pages 87–91, Gaithersburg, Maryland, USA.

Bernstein, J. and Pisoni, D. B. (1980). Unlimited Text-to-Speech System: Description and Evaluation of a Microprocessor Based Device. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 576–579, Denver, Colorado, USA.

Bladon, A. (1990). Evaluating the Prosody of Text-to-Speech Synthesizers. In *Proceedings of Speech Tech*, pages 215–220, New York, USA.

Boeffard, O., Cherbonnel, B., Emerard, F., and White, S. (1993). Automatic Segmentation and Quality Evaluation of Speech Unit Inventories for Concatenation-Based, Multilingual PSOLA Text-to-Speech Systems. In *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech)*, volume 2, pages 1449–1452, Berlin, Germany.

Boogaar, T. and Silverman, K. (1992). Evaluating the Overall Comprehensibility of Speech Synthesizers. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, pages 1207–1210, Banff, Canada.

Calliope (1989). Évaluation de la parole codée et synthétique. In *La parole et son traitement automatique, Collection scientifique et technique des Télécommunications*, pages 455–488, Masson, Paris, France.

Carlson, R. and Granström, B. (1989). Evaluation and Development of the KTH Text-to-Speech System on the Segmental Level. In *Proceedings of the ESCA Tutorial and Research Workshop on Speech Input/Output Assess-*

ment and Speech Databases, volume 2, pages 11–14, Noordwijkerhout, The Netherlands.

Carlson, R., Granström, B., and Larson, K. (1976). Evaluation of a Text-to-Speech System as a Reading Machine for the Blind. In *STL-QPSR 2-3, Quarterly Progress and Status Report*, pages 9–13, KTH, Stockholm, Sweden.

Carlson, R., Granström, B., Neovius, L., and Nord, L. (1992). The 'Listening Speed' Paradigm for Synthesis Evaluation. In *FONETIK, Sixth Swedish Phonetics Conference, Chalmers Technical Report No. 10*, pages 63–66, Gothenburg, Sweden. Department of Information Theory, Chalmers University of Technology.

Carlson, R., Granström, B., and Nord, L. (1990a). Results from the SAM Segmental Test for Synthetic and Natural Speech in Swedish (VCV, CV and VC). Internal Report, ESPRIT Project 2589 (SAM) Multi-Lingual Speech Input/Output, Assessment, Methodology and Standardization.

Carlson, R., Granström, B., and Nord, L. (1990b). Segmental Intelligibility of Synthetic and Natural Speech in Real and Nonsense Words. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, volume 2, pages 989–992, Kobe, Japan.

Cartier, M. and Gleiss, N. (1992). Synthetic Speech Quality Assessment for Telecommunication Purposes. In *Proceedings of the COCOSDA Meeting*, Banff, Canada.

Cartier, M., Karlsson, C., and Modena, G. (1989). Standardization of Synthetic Speech Quality for Telecommunication Purposes. In *Proceedings of the ESCA Tutorial and Research Workshop on Speech Input/Output Assessment and Speech Databases*, volume 2, pages 99–102, Noordwijkerhout, The Netherlands.

CCITT (1992a). *A Method for Subjective Performance Assessment of the Quality of Speech of Voice Output Devices*. Comité Consultatif International Téléphonique et Télégraphique, Draft Recommendation P.8S of Working Party XII/2, Special Rapporteur for Question 5/XII edition. Available upon request. International Telecommunications Union, Geneva, Switzerland.

CCITT (1992b). *Experiment in Assessing the Quality of Synthetic Speech*. Comité Consultatif International Téléphonique et Télégraphique, Temporary Document No. 70-E of Working Party XII/2 edition. Special Rapporteur for Question 5/XII. International Telecommunications Union, Geneva, Switzerland.

Delogu, C., Conte, S., Paoloni, A., and Sementina, C. (1992a). Comprehension of Synthetic Speech in Good and in Adverse Conditions. In *Proceedings of the ESCA Workshop on Speech Processing in Adverse Conditions*, pages 53–56, Cannes-Mandelieu, France.

Delogu, C., Conte, S., Paoloni, A., and Sementina, C. (1992b). Two Different Methodologies for Evaluating the Comprehension of Synthetic Passages. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, volume 2, pages 1231–1234, Banff, Canada.

Delogu, C., Di Carlo, A., Sementina, C., and Stecconi, S. (1993a). A Methodology for Evaluating Human-Machine Spoken Language Interaction. In *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech)*, pages 1427–1430, Berlin, Germany.

Delogu, C., Falcone, M., Paoloni, A., Ridolfi, P., and Vagges, K. (1992c). Intelligibility of Italian Text-to-Speech Synthesizers in Adverse Conditions. In *Proceedings of the ESCA Workshop on Speech Processing in Adverse Conditions*, pages 57–60, Cannes-Mandelieu, France.

Delogu, C., Paoloni, A., and Pocci, P. (1991a). New Directions in the Evaluation of Voice Input/Output Systems. *IEEE Journal on Selected Areas in Communications*, 9(4):566–573.

Delogu, C., Paoloni, A., Pocci, P., and Sementina, C. (1991b). Quality Evaluation of Text-to-Speech Synthesizers Using Magnitude Estimation, Categorical Estimation, Pair Comparison, and Reaction Time Methods. In *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech)*, pages 353–356, Genova, Italy.

Delogu, C., Paoloni, A., Ridolfi, P., and Vagges, K. (1993b). Intelligibility of Speech Produced by Text-to-Speech Synthesizers Over the Orthophonic and Telephonic Channel. In *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech)*, pages 1893–1896, Berlin, Germany.

Delogu, C. and Sementina, C. (1993). Towards a More Realistic Evaluation of Synthetic Speech: A Cognitive Perspective. In *Proceedings of the NATO-ASI Workshop on New Advances and Trends in Speech Recognition and Coding*, Bubion, Granada, Spain.

Falaschi, A. (1992). Segmental Quality Assessment by Pseudo-Words. In Bailly, G., Benoit, C., and Sawallis, T., editors, *Talking Machines: Theories, Models and Designs*, pages 455–472, Elsevier, North Holland, Amsterdam, The Netherlands.

Fourcin, A. (1992). Assessment of Synthetic Speech. In Bailly, G., Benoit, C., and Sawallis, T., editors, *Talking Machines: Theories, Models and Designs*, pages 431–434, Elsevier, North Holland, Amsterdam, The Netherlands.

Fourcin, A. J., Harland, G., Barry, W., and Hazan, V., editors (1989). *Speech Input and Output Assessment. Multilingual Methods and Standards*. Ellis Horwood Ltd., Chichester, UK.

French PTT (1987). *An 'Objective' Evaluation of Difficulty in Understanding Voice Synthesis Devices*. Comité Consultatif International Téléphonique et Télégraphique, International Telecommunications Union, Geneva,

Switzerland. CCITT Com. XII, Study Group XII, Contribution No. FR4-E.

Garnier-Rizet, M. (1993). Evaluation of a Rule-Based Text-to-Speech System for French at the Segmental Level. In *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech)*, pages 1889–1892, Berlin, Germany.

Goldstein, M., Lindström, B., and Till, O. (1992). Assessing Global Performance of Speech Synthesizers: Context Effects When Assessing Naturalness of Swedish Sentence Pairs Generated by 4 Systems Using 3 Different Assessment Procedures (Free Number Magnitude Estimation, 5-and 11-Point Category Scales). ESPRIT Project 2589 (SAM), Multilingual Speech Input/Output Assessment, Methodology and Standardization. Part of SAM Final Report, University College London, UK, 19 pages.

Graillot, P. (1983). Synthèse de la parole: Évaluation de la qualité en recette. I. S. F., Le Traitement Automatique de la Parole, CNET (French Telecom), Paris, France.

Graillot, P., Emerard, F., and Le Bras, J. (1983). Tests de rime appliqués a des systèmes de synthèse par diphones. Note Technique CNET, NT/LAA/TSS/180, CNET (French Telecom), 55 pages.

Granström, B. and Nord, L. (1989). A Report on Swedish Phonotactic Structures and Constraints. Interim Report, Year One, ESPRIT Project 2589 (SAM) Multi-Lingual Speech Input/Output, Assessment, Methodology and Standardization, pages 135–156.

Greene, B. G., Manous, L. M., and Pisoni, D. B. (1984). Perceptual Evaluation of DECtalk: A First Report on Perfect Paul. Speech Research Laboratory Technical Note No. 84-03, Speech Research Laboratory, Indiana University, Bloomington, USA.

Hampshire, B., Ruden, J., Carlson, R., and Granström, B. (1982). Evaluation of Centrally Produced and Distributed Synthetic Speech. In *STL-QPSR 2-3, Quarterly Progress and Status Report*, pages 18–23, KTH, Stockholm, Sweden.

House, J., MacDermid, C., McGlashan, S., Simpson, A., and Youd, N. J. (1993). Evaluating Synthesised Prosody in Simulations of an Automated Telephone Enquiry Service. In *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech)*, pages 901–904, Berlin, Germany.

House, J. and Youd, N. J. (1992). Evaluating the Prosody of Synthesised Utterances within a Dialogue System. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, pages 1175–1178, Banff, Canada.

Houtgast, T. and Steeneken, H. J. M. (1971). Evaluation of Speech Transmission Channels by Using Artificial Signals. *Acustica*, 25:355–367.

Houtgast, T. and Steeneken, H. J. M. (1984). A Multi-Lingual Evaluation of the Rasti-Method for Estimating Speech Intelligibility in Auditoria. *Acustica*, 54(1):85–199.

Houtgast, T. and Verhave, J. A. (1991). A Physical Approach to Speech Quality Assessment: Correlation Patterns in the Speech Spectrogram. In *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech)*, volume 1, pages 285–288, Genova, Italy.

Houtgast, T. and Verhave, J. A. (1992). An Objective Approach to Speech Quality. ESPRIT Project 2589 (SAM), Multilingual Speech Input/Output Assessment, Methodology and Standardization. Part of SAM Final Report, University College London, UK, 17 pages.

Howard-Jones, P. (1992). SOAP, Speech Output Assessment Package, Version 4.0. ESPRIT Project 2589 (SAM) Report, University College London, UK.

Jekosch, U. (1989). The Cluster-Based Rhyme Test: A Segmental Synthesis Test For Open Vocabulary. In *Proceedings of the ESCA Tutorial and Research Workshop on Speech Input/Output Assessment and Speech Databases*, volume 2, pages 15–18, Noordwijkerhout, The Netherlands.

Jekosch, U. (1990). A Weighted Intelligibility Measure for Speech Assessment. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, volume 2, pages 973–976, Kobe, Japan.

Jekosch, U. (1992). The Cluster-Identification Test. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, volume I, pages 205–209, Banff, Canada.

Jekosch, U. (1993). Cluster-Similarity: A Useful Database for Speech Processing. In *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech)*, pages 195–198, Berlin, Germany.

Karlsson, I. (1992). Evaluations of Acoustic Differences Between Male and Female Voices: A Pilot Study. In *STL-QPSR 1, Quarterly Progress and Status Report*, pages 19–31, KTH, Stockholm, Sweden.

Kasuya, H. (1993). Significance of Suitability Assessment in Speech Synthesis Applications. *IEICE Transactions on Fundamentals of Electronics Communications and Computer Sciences*, E76-A(11):1893–1897.

Kasuya, H. and Kasuya, S. (1992). Relationships between Syllable, Word and Sentence Intelligibilities of Synthetic Speech. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, pages 1215–1218, Banff, Canada.

Kitawaki, N. (1991). Quality Assessment of Coded Speech. In Furui, S. and Sondhi, M. M., editors, *Advances in Speech Signal Processing*, chapter 12, pages 357–385, Marcel Dekker, New York, USA.

Kitawaki, N. and Nagabuchi, H. (1988). Quality Assessment of Speech Coding and Speech Synthesis Systems. *IEEE Communications Magazine*, 26(10):36–44.

Klaus, H., Klix, H., Sotscheck, J., and Fellbaum, K. (1993). An Evaluation System for Ascertaining the Quality of Synthetic Speech Based on Subjective Category Rating Tests. In *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech)*, pages 1679–1682, Berlin, Germany.

Logan, J. S. and Greene, B. G. (1985). Perceptual Evaluation of DECtalk V1.8: Identification of Isolated Phonetically Balanced (PB) Words. Speech Research Laboratory Technical Note No. 85-04, Speech Research Laboratory, Indiana University, Bloomington, USA.

Logan, J. S., Greene, B. G., and Pisoni, D. B. (1985a). Perceptual Evaluation of the Prose 3.0. Text-to-Speech System: Phoneme Intelligibility Measured Using the MRT. Speech Research Laboratory Technical Note No. 85-05, Speech Research Laboratory, Indiana University, Bloomington, USA.

Logan, J. S., Greene, B. G., and Pisoni, D. B. (1989a). Measuring the Segmental Intelligibility of Synthetic Speech Produced by Ten Text-to-Speech Systems. *Journal of the Acoustical Society of America*, 86:566–581.

Logan, J. S., Greene, B. G., and Pisoni, D. B. (1989b). Segmental Intelligibility of Synthetic Speech Produced by Rule. *Journal of the Acoustical Society of America*, 86(2):566–581.

Logan, J. S. and Pisoni, D. B. (1986a). Intelligibility of Phoneme Specific Sentences Using Three Text-to-Speech Systems and a Natural Speech Control. In *Research on Speech Perception, Progress Report 12*, pages 319–333, Speech Research Laboratory, Indiana University, Bloomington, USA.

Logan, J. S. and Pisoni, D. B. (1986b). Preference Judgements Comparing Different Synthetic Voices. In *Research on Speech Perception, Progress Report 12*, pages 263–289, Speech Research Laboratory, Indiana University, Bloomington, USA.

Logan, J. S., Pisoni, D. B., and Greene, B. G. (1985b). Measuring the Segmental Intelligibility of Synthetic Speech: Results from Eight Text-to-Speech Systems. In *Research on Speech Perception, Progress Report 11*, pages 3–31, Speech Research Laboratory, Indiana University, Bloomington, USA.

Lopez-Gonzalo, E., Olaszy, G., and Nemeth, G. (1993). Improvements of the Spanish Version of the Multivox Text-to-Speech System. In *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech)*, pages 869–872, Berlin, Germany.

Manous, L. M., Greene, B. G., and Pisoni, D. B. (1984). Perceptual Evaluation of Prose - the Speech Plus Text-to-Speech System: I. Phoneme Intelligibility and Word Recognition in Meaningful Sentences. Speech Research Laboratory Technical Note 84-04, Speech Research Laboratory, Indiana University, Bloomington, USA.

Manous, L. M. and Pisoni, D. B. (1984). Effects of Signal Duration on the Perception of Natural and Synthetic Speech. In *Research on Speech Perception,*

Progress Report 10, pages 311–321, Speech Research Laboratory, Indiana University, Bloomington, USA.

Manous, L. M., Pisoni, D. B., Dedina, M. J., and Nusbaum, H. C. (1985). Comprehension of Natural and Synthetic Speech Using a Sentence Verification Task. In *Research on Speech Perception, Progress Report 11*, pages 33–57, Speech Research Laboratory, Indiana University, Bloomington, USA.

Mariniak, A. (1993). A Global Framework for the Assessment of Synthetic Speech without Subjects. In *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech)*, pages 1683–1686, Berlin, Germany.

Monaghan, A. I. C. (1989). Evaluating Intonation in the CSTR Text-to-Speech System. In *Proceedings of the ESCA Tutorial and Research Workshop on Speech Input/Output Assessment and Speech Databases*, volume 2, pages 111–114, Noordwijkerhout, the Netherlands.

Monaghan, A. I. C. and Ladd, D. R. (1990). Symbolic Output as the Basis for Evaluating Intonation in Text-to-Speech Systems. *Speech Communication*, 9(4):305–314.

Nusbaum, H. C., Dedina, M. J., and Pisoni, D. B. (1984a). Perceptual Confusions of Consonants in Natural and Synthetic CV Syllables. Speech Research Laboratory Technical Note 84-02, Speech Research Laboratory, Indiana University, Bloomington, USA.

Nusbaum, H. C., Greenspan, S. L., and Pisoni, D. B. (1986). Perceptual Attention in Monitoring Natural and Synthetic Speech. In *Research on Speech Perception, Progress Report 12*, pages 307–318, Speech Research Laboratory, Indiana University, Bloomington, USA.

Nusbaum, H. C. and Pisoni, D. B. (1984). Perceptual Evaluation of Synthetic Speech Generated by Rule. In *Proceedings of the Fourth Voice Data Entry Systems Applications Conference*, Sunnyvale, California, USA.

Nusbaum, H. C., Pisoni, D. B., Schwab, E. C., Luce, P. A., and Slowiaczek, L. M. (1983a). Perception of Synthetic Speech under high Cognitive Load. Paper presented at the Voice SubTAG Meeting, Fort Monmouth, New Jersey, USA.

Nusbaum, H. C., Schwab, E. C., and Pisoni, D. B. (1983b). Perceptual Evaluation of Synthetic Speech: Some Constraints on the Use of Voice Response Systems. In *Research on Speech Perception, Progress Report 9*, pages 283–294, Speech Research Laboratory, Indiana University, Bloomington, USA. Also published in *Proceedings of the Voice Data Entry Systems Application Conference,* pages 1–14, Chicago, USA.

Nusbaum, H. C., Schwab, E. C., and Pisoni, D. B. (1984b). Subjective Evaluation of Synthetic Speech: Measuring Preference, Naturalness and Acceptibility. In *Research on Speech Perception, Progress Report 10*, pages

391–407, Speech Research Laboratory, Indiana University, Bloomington, USA.

Nye, P. W. and Gaitenby, J. H. (1973). Consonant Intelligibility in Synthetic Speech and in Natural Speech Control (Modified Rhyme Test Results). In *Haskins Laboratories Status Report on Speech Research, SR-33*, pages 77–91.

Nye, P. W. and Gaitenby, J. H. (1974). The Intelligibility of Synthetic Monosyllabic Words in Short Syntactically Normal Sentences. In *Haskins Laboratories Status Report on Speech Research, SR-37/38*, pages 169–190.

Nye, P. W., Ingemann, F., and Donald, L. (1975). Synthetic Speech Comprehension: A Comparison of Listener Performances with and Preferences among Different Forms. In *Haskins Laboratories Status Report on Speech Research, SR-41*, pages 117–126.

O'Malley, M. H. and Caisse, M. (1987). How to Evaluate Text-to-Speech Systems. *Speech Technology*, 3(4):66–75.

Ozawa, K. and Logan, J. S. (1989). Perceptual Evaluation of Two Speech Coding Methods by Native and Non-native Speakers of English. *Computer Speech and Language*, 3:53–59.

Pallett, D. H., editor (1982). *Proceedings of the NBS Workshop on Standardization for Speech I/O Technology*, Gaithersburg, Maryland, USA.

Pascal, D. (1987). Méthodologies d'évaluation subjective de la qualité des systèmes de communication. In *Dossier: La qualité des services de communication*. Bulletin No. 28 de l'Institut pour le Développement et l'Amenagement des Télécommunications et de l'Économie, Montpellier, France.

Pascal, D. and Combescure, P. (1988). Évaluation de la qualité de la transmission vocale. L'Écho des Recherches, 132: 31–40.

Pavlovic, C. V. (1987). Derivation of Primary Parameters and Procedures for Use in Speech Intelligibility Predictions. *Journal of the Acoustical Society of America*, 82:413–422.

Pavlovic, C. V. and Rossi, M. (1990). Quality Assessment of Synthesized Speech: Status Report, Systematization, and Recommendations. In *Esprit Project 2589 (SAM), Interim Report, Year One*, pages 354–361.

Pavlovic, C. V., Rossi, M., and Espesser, R. (1989a). A Pilot Study on the Possibility of Using the ESNR Method for Assessing Text-to-Speech Synthesis Systems. In *Final Report, Extension Phase, ESPRIT Project 1542 (SAM)*, pages 40–42.

Pavlovic, C. V., Rossi, M., and Espesser, R. (1989b). Direct Scaling of the Performance of Text-to-Speech Synthesis Systems. In *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech)*, volume 2, pages 644–647, Paris, France.

Pavlovic, C. V., Rossi, M., and Espesser, R. (1989c). Subjective Assessment of Acceptability, Intelligibility and Naturalness of Text-to-Speech Synthesis. In *Proceedings of the ESCA Tutorial and Research Workshop on Speech Input/Output Assessment and Speech Databases*, volume 2, pages 94–98, Noordwijkerhout, The Netherlands.

Pavlovic, C. V., Rossi, M., and Espesser, R. (1990). Use of the Magnitude Estimation Technique for Assessing the Performance of Text-to-Speech Synthesis Systems. In *Interim Report, Year One, Esprit Project 2589 (SAM)*, pages 362–380.

Pavlovic, C. V., Rossi, M., and Espesser, R. (1991a). Definition of Assessment Methodology for Overall Quality of Synthetic Speech. Stage Report 3, Year 2, Interim Report, SAM-UCL-G003, ESPRIT Project 2589 (SAM).

Pavlovic, C. V., Rossi, M., and Espesser, R. (1991b). Methods for Reducing Context Effects in the Subjective Assessment of Synthetic Speech. In *Proceedings of the 12th International Congress on Phonetic Sciences*, pages 82–85, Aix-en-Provence, France.

Picone, J., Goudie-Marshall, K. M., Doddington, G. R., and Fisher, W. (1986). Automatic Text Alignment for Speech Systems Evaluation. *IEEE Transactions on Acoustics, Speech, and Signal Processing (ASSP)*, 34(4):780–784.

Pisoni, D. B. (1989). Perceptual Evaluation of Synthetic Speech: A Tutorial Review. In *Proceedings of the ESCA Tutorial and Research Workshop on Speech Input/Output Assessment and Speech Databases*, Noordwijkerhout, The Netherlands, 13 pages.

Pisoni, D. B. and Dedina, M. J. (1986). Comprehension of Digitally Encoded Natural Speech Using a Sentence Verification Task (SVT): A First Report. In *Research on Speech Perception, Progress Report 12*, pages 3–18, Speech Research Laboratory, Indiana University, Bloomington, USA.

Pisoni, D. B. and Greene, B. G. (1990). The Role of Cognitive Factors in the Perception of Synthetic Speech. In Fujisaki, H., editor, *International Symposium on International Coordination and Standardization of Speech Database and Assessment Techniques for Speech Input/Output*, pages 3–25, Kobe, Japan.

Pisoni, D. B., Greene, B. G., and Logan, J. S. (1989). An Overview of Ten Years of Research on the Perception of Synthetic Speech. In *Proceedings of the ESCA Tutorial and Research Workshop on Speech Input/Output Assessment and Speech Databases*, volume 2, pages 1–4, Noordwijkerhout, The Netherlands.

Pisoni, D. B., Greene, B. G., and Nusbaum, H. C. (1985a). Some Human Factors Issues in the Perception of Synthetic Speech. In *Proceedings of Speech Tech*, pages 57–61, New York, USA.

Pisoni, D. B. and Hunnicutt, S. (1980). Perceptual Evaluation of MITalk: The MIT Unrestricted Text-to-Speech System. In *Proceedings of the Interna-*

tional Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 572–575, Denver, Colorado, USA.

Pisoni, D. B. and Koen, E. (1981). Some Comparison of Intelligibility of Synthetic and Natural Speech at Different Speech-to-Noise Ratios. In *Research on Speech Perception, Progress Report 7*, pages 243–253, Speech Research Laboratory, Indiana University, Bloomington, USA.

Pisoni, D. B., Manous, L. M., and Dedina, M. J. (1986). Comprehension of Natural and Synthetic Speech: II. Effects of Predictability on Verification of Sentences Controlled for Intelligibility. In *Research on Speech Perception, Progress Report 12*, pages 19–41, Speech Research Laboratory, Indiana University, Bloomington, USA.

Pisoni, D. B., Nusbaum, H., and Greene, B. G. (1985b). Perception of Synthetic Speech Generated by Rule. In *Proceedings of the Institute of Electrical and Electronics Engineers (IEEE)*, 73(11): 1665–1676.

Pisoni, D. B. and Nusbaum, H. C. (1986). Developing Methods for Assessing the Performance of Speech Synthesis and Recognition Systems. In *Proceedings of the Human Factors Society*, volume 2, pages 1344–1348, Santa Monica, California, USA. Human Factors Society. (Also appeared in Speech Research Laboratory Publications and Reports 12, Indiana University, Bloomington, USA, 1986.).

Pisoni, D. B., Nusbaum, H. C., Luce, P. A., and Schwab, E. C. (1983). Perceptual Evaluation of Synthetic Speech: Some Considerations of the User/System Interface. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 535–538, Boston, USA.

Pisoni, D. B., Ralston, J. V., and Lively, S. E. (1990). Some New Directions in Research on Comprehension of Synthetic Speech. In Fujisaki, H., editor, *International Symposium on International Coordination and Standardization of Speech Database and Assessment Techniques for Speech Input/Output*, pages 29–42, Kobe, Japan.

Plomp, R. and Mimpen, A. M. (1979). Improving the Reliability of Testing the Speech Reception Threshold for Sentences. *Audiology*, 8:43–52.

Pols, L. C. W. (1974). Intelligibility of Speech Resynthesized by Using a Dimensional Spectral Representation. In *Speech Communication Seminar*, pages 87–95, Stockholm, Sweden.

Pols, L. C. W. (1975). Analysis and Synthesis of Speech Using a Broad-Band Spectral Representation. In Fant, G. and Tatham, M. A., editors, *Symposium on Auditory Analysis and Perception of Speech*, pages 23–36, Leningrad, Russia. Academic Press.

Pols, L. C. W. (1979a). Consonant Intelligibility of Speech Produced by Dyadic Rule Synthesis. British Telecom Laboratories Technical Memorandum, TM-79-1228-5, 21 pages.

Pols, L. C. W. (1979b). Intelligibility of Intervocalic Consonants in Noise. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 460–463, Washington DC, USA. Also IZF Report 1978-C25 (expanded version).

Pols, L. C. W. (1986). Assessing the Performance of Speech Technology Systems. In Clark, J. E., editor, *An Introduction to Speech Science*, chapter 5, pages 18–20, First Australian Conference on Speech Science and Technology, Canberra, Australia.

Pols, L. C. W. (1987). Quality Evaluation of Text-to-Speech Synthesis Systems. In *Multilingual Speech Input-Output Assessment, Methodology, and Standardization*. Deliverable of ESPRIT-project 1541. Also IFA Report No. 94, 31 pages.

Pols, L. C. W. (1988). Improving Synthetic Speech Quality by Systematic Evaluation. In *Proceedings of the Institute of Phonetic Sciences, Amsterdam (IFA) 12*, pages 19–27, Amsterdam, The Netherlands. Also in Preprints of the Second Symposium on Advanced Man-Machine Interface Through Spoken Language, November, Makaha, Oahu, Hawaii, USA, 17.1–17.9.

Pols, L. C. W. (1989a). Assessment of Text-to-Speech Synthesis Systems. In Fourcin, A., Harland, G., Barry, W., and Hazan, V., editors, *Speech Input and Output Assessment. Multilingual Methods and Standards*, chapter III, pages 53–81 and 251–266, Ellis Horwoord, Chichester, UK.

Pols, L. C. W. (1989b). ESCA Tutorial Day and Workshop on Speech Input/Output Assessment and Speech Databases. *Speech Communication*, 8(4):379–380.

Pols, L. C. W. (1990a). Assessing the Speech Quality of Text-to-Speech Synthesizers. In *Proceedings of VERBA, International Conference on Speech Technology*, pages 295–302, Rome, Italy.

Pols, L. C. W. (1990b). Does Improved Performance also Contribute to more Phonetic Knowledge? In *Proceedings of the ESCA Tutorial Day on Speech Synthesis*, pages 49–54, Autrans, France.

Pols, L. C. W. (1990c). How Useful are Speech Databases for Rule Synthesis Development and Assessment? In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, volume 2, pages 1289–1292, Kobe, Japan.

Pols, L. C. W. (1990d). Improving Synthetic Speech Quality by Systematic Evaluation. In Fujisaki, H., editor, *Recent Research Toward Advanced Man-Machine Interface Through Spoken Language*, pages 445–453, Tokyo, Japan.

Pols, L. C. W. (1990e). Special Issue on Speech Input/Output Assessment and Speech Databases. *Speech Communication*, 9(4):263–388.

Pols, L. C. W. (1990f). 'Standardized' Synthesis Evaluation Methods. In *Proceedings of the International Workshop on International Coordination and*

Standardization of Speech Database and Assessment Techniques for Speech Input/Output, pages 53–60, Kobe, Japan.

Pols, L. C. W. (1991a). Evaluating the Performance of Speech Input/Output Systems. A Report on the ESPRIT-SAM Project. In *Proceedings of Tagung der Deutschen Arbeitsgemeinschaft für Akustik (DAGA)*, pages 139–150, Bochum, Germany.

Pols, L. C. W. (1991b). Evaluating the Performance of Speech Technology Systems. In *Proceedings of the Institute of Phonetic Sciences, Amsterdam (IFA) 15*, pages 27–41, Amsterdam, The Netherlands.

Pols, L. C. W. (1991c). Quality Assessment of Text-to-Speech Synthesis-by-Rule. In Furui, S. and Sondhi, M. M., editors, *Advances in Speech Signal Processing*, chapter 13, pages 387–416, Marcel Dekker, New York, USA.

Pols, L. C. W. (1994). Speech Technology Systems: Performance and Evaluation. In *Encyclopedia of Language & Linguistics*, volume 8, pages 4289–4296, Pergamon Press, Oxford, UK.

Pols, L. C. W. and Boxelaar, G. W. (1986). Comparative Evaluation of the Speech Quality of Speech Coders and Text-to-Speech Synthesizers. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 901–904, Tokyo, Japan.

Pols, L. C. W., Lefevre, J. P., Boxelaar, G., and van Son, N. (1987). Word Intelligibility of a Rule Synthesis System for French. In *Proceedings of the European Conference on Speech Technology*, volume 1, pages 179–182, Edinburgh, UK.

Pols, L. C. W. and Olive, J. P. (1983). Intelligibility of Consonants in CVC Utterances Produced by Dyadic Rule Synthesis. *Speech Communication*, 2(1):3–13.

Pols, L. C. W. and SAM-partners (1992). Multi-Lingual Synthesis Evaluation Methods. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, volume 1, pages 181–184, Banff, Canada.

Pols, L. C. W. and van Bezooijen, R. (1991). Gaining Phonetic Knowledge whilst Improving Synthetic Speech Quality? *Journal of Phonetics*, 19(1):139–146.

Pols, L. C. W. and van Rooijen, J. N. M. (1974). Intelligibility of Speech Resynthesized by Using a Dimensional Spectral Representation. In *Proceedings of the Eighth International Conference on Acoustics (ICA)*, page 297, London, UK.

Portele, T. (1993). Evaluation der segmentalen Verständlichkeit des Sprachsynthesesystems HADIFIX mit der SAM-Testprozedur. In *Proceedings of Tagung der Deutschen Arbeitsgemeinschaft für Akustik (DAGA)*, pages 1032–1035, Frankfurt, Germany.

Pratt, R. L. (1984). The Assessment of Speech Intelligibility at RSRE. In *Proceedings of the Instititute of Acoustics*, volume 6, Part 4, pages 439–443.

Pratt, R. L. (1986). On the Intelligibility of Synthetic Speech. In *Proceedings of the Instititute of Acoustics*, volume 8, Part 7, pages 183–192.

Pratt, R. L. (1987). Quantifying the Performance of Text-to-Speech Synthesizers. *Speech Technology*, 3(4):54–64.

Purdy, S. C. and Pavlovic, C. V. (1991). Scaling of Speech Intelligibility Using Magnitude Estimation and Category Estimation and Paired Comparisons. In *Proceedings of the 12th International Congress on Phonetic Sciences Sciences*, pages 434–437, Aix-en-Provence, France.

Purdy, S. C. and Pavlovic, C. V. (1992). Reliability, Sensitivity and Validity of Magnitude Estimation, Category Scaling and Paired-Comparison Judgments of Speech Intelligibility by Older Listeners. *Audiology*, 31:254–271.

Robert, J.-M., Choiniere, A., and Descout, R. (1989). Subjective Evaluation of Naturalness and Acceptability of Three Text-to-Speech Systems in French. In *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech)*, volume 2, pages 640–643, Paris, France.

Rossi, M., Espesser, R., and Pavlovic, C. V. (1991a). Subjective Rating of Text-to-Speech Synthesis: The Effects of an Internal Reference and Cross-Modality Matching. *VERBUM, Tome XIV, fascicule 2-3-4*, pages 367–377.

Rossi, M., Espesser, R., and Pavlovic, C. V. (1991b). The Effects of an Internal Reference System and Cross-Modality Matching on the Subjective Rating of Speech Synthesisers. In *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech)*, volume 1, pages 273–276, Genova, Italy.

Rossi, M., Pavlovic, C., and Espesser, R. (1989). Magnitude Estimation Technique in Evaluating Text-to-Speech Synthesis. In *Proceedings of the 13th International Congress on Acoustics*, volume 2, pages 479–482, Belgrade, Yugoslavia.

Rosson, M. B. (1985). Listener Training for Speech-Output Applications. In Borman, L. and Curtis, B., editors, *Proceedings of the Conference on Human Factors in Computing Systems*, pages 193–196, New York, USA. ACM.

Salza, P. L., Sandri, S., and Foti, E. (1987). Evaluation of Experimental Diphones for Text-to-Speech Synthesis. In *Proceedings of the European Conference on Speech Technology*, volume 1, pages 63–66, Edinburgh, UK.

Schnabel, B. (1986). Évaluation de la qualité de l'allemand synthétisé par diphones. In *15èmes Journées d'Etudes sur la Parole, GALF*, pages 19–20, Aix-en-Provence, France.

Schwab, E. C., Nusbaum, H. C., and Pisoni, D. B. (1983). Some Effects of Training on the Perception of Synthetic Speech. In *Research on Speech Perception, Progress Report 9*, pages 39–77, Indiana University, USA.

Schwab, E. C. and Pisoni, D. B. (1983). The Effects of Training on the Intelligibility of Synthetic Speech: I. Pre-Test and Post-Test Data. *Journal of the Acoustical Society of America*, 73, S3 (A).

Silverman, K., Basson, S., and Levas, S. (1990). Evaluating Synthesis Performance: Is Segmental Intelligibility Enough? In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, pages 981–984, Kobe, Japan.

Sorin, C. (1982/83). Évaluation de la contribution de F0 a l'intelligibilité. *Recherches Acoustiques, CNET*, 7:141–155.

Sorin, C., Benoit, C., Charpentier, F., Emerard, F., and Schnabel, B. (1988). Évaluation de systèmes de synthèse. Contribution CNET au Projet Esprit SAM-Speech Assessment Methods, France.

Spiegel, M. F., Altom, M. J., Macchi, M. J., and Wallace, K. L. (1990). Comprehensive Assessment of the Telephone Intelligibility of Synthesized and Natural Speech. *Speech Communications*, 9:279–292.

Steeneken, H. J. M. (1982). Ontwikkeling en toetsing van een Nederlandstalige diagnostische rijmtest voor het testen van spraakkommunikatiekanalen. IZF Report 1982-13, TNO Institute for Perception, Soesterberg, The Netherlands (in Dutch).

Steeneken, H. J. M. (1987a). Comparison among Three Subjective and One Objective Intelligibility Test. IZF Report 1987-8, TNO Institute for Perception, Soesterberg, The Netherlands.

Steeneken, H. J. M. (1987b). Diagnostic Information of Subjective Intelligibility Tests. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 131–134, Dallas, USA.

Steeneken, H. J. M. (1992a). *On Measuring and Predicting Speech Intelligibility*. Doctoral thesis, University of Amsterdam, The Netherlands.

Steeneken, H. J. M. (1992b). Quality Evaluation of Speech Processing Systems. In Ince, N., editor, *Digital Speech Coding: Speech Coding, Synthesis and Recognition*, chapter 5, pages 127–160, Kluwer, Norwell, USA.

Steeneken, H. J. M. and Geurtsen, F. W. M. (1988). Description of the RSG-10 Noise Data-Base. IZF Report 1988-3, TNO Institute for Perception, Soesterberg, The Netherlands.

Steeneken, H. J. M., Geurtsen, F. W. M., and Agterhuis, E. (1990). Speech Data-Base for Intelligibility and Speech Quality Measurements. IZF Report 1990 A-13, TNO Institute for Perception, Soesterberg, The Netherlands.

Steeneken, H. J. M. and Houtgast, T. (1973). Intelligibility in Telecommunication Derived from Physical Measurements. In *Proceedings of the Symposium on Intelligibilité de la Parole*, pages 73–80, Liège, Belgium.

Steeneken, H. J. M. and Houtgast, T. (1978). Comparison of some Methods for Measuring Speech Levels. IZF Report 1978-22, TNO Institute for Perception, Soesterberg, The Netherlands.

Steeneken, H. J. M. and Houtgast, T. (1980). A Physical Method for Measuring Speech-Transmission Quality. *Journal of the Acoustical Society of America*, 67:318–326.

Steeneken, H. J. M. and Houtgast, T. (1991). On the Mutual Dependency of Octave-Band Specific Contributions to Speech Intelligibility. In *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech)*, pages 1133–1136, Genova, Italy.

Steeneken, H. J. M. and van Velden, J. G. (1991). RAMOS Recognizer Assessment by Means of Manipulation of Speech Applied to Connected Speech Recognition. In *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech)*, pages 529–532, Genova, Italy.

Steeneken, H. J. M., Verhave, J. A., and Houtgast, T. (1993). Objective Assessment of Speech Communication Systems. In *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech)*, pages 203–206, Berlin, Germany.

Sydeserff, H. A., Caley, R. J., Isard, S. D., Jack, M. A., Monaghan, A. I. C., and Verhoeven, J. (1991). Evaluation of Synthetic Speech Techniques in a Comprehension Task. In *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech)*, volume 1, pages 277–280, Genova, Italy.

Syrdal, A. K. (1987). Methods for a Detailed Analysis of Dynastat DRT Results. Technical Memorandum, AT&T Bell Laboratories.

Syrdal, A. K. and Sciacca, B. A. (1993). A Diagnostic Text-to-Speech Intelligibility Test. *Journal of the Acoustical Society of America*, 94:paper 4aSP9, 1842–1843 (A).

van Bezooijen, R. (1988). Evaluation of the Quality of Consonant Clusters in Two Synthesis Systems for Dutch. In *Proceedings of Speech'88, Seventh FASE Symposium*, volume 2, pages 445–452, Edinburgh, UK.

van Bezooijen, R. and Jongenburger, W. (1993). Evaluation of an Electronic Newspaper for the Blind in the Netherlands. In Granström, B., Hunnicutt, S., and Spens, E., editors, *Proceedings of the ESCA Workshop on Speech and Language Technology for Disabled Persons*, pages 195–198, Stockholm, Sweden.

van Bezooijen, R. and Pols, L. C. W. (1987). Evaluation of Two Synthesis-by-Rule Systems for Dutch. In *Proceedings of the European Conference on Speech Technology*, volume 1, pages 183–186, Edinburgh, UK.

van Bezooijen, R. and Pols, L. C. W. (1989a). Evaluation of a Sentence Accentuation Algorithm for a Dutch Text-to-Speech System. In *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech)*, volume 1, pages 218–221, Paris, France.

van Bezooijen, R. and Pols, L. C. W. (1989b). Evaluation of Text-to-Speech Conversion for Dutch: From Segment to Text. In *Proceedings of the ESCA Tutorial and Research Workshop on Speech Input/Output Assessment and Speech Databases*, volume 2, pages 103–106, Noordwijkerhout, The Netherlands.

van Bezooijen, R. and Pols, L. C. W. (1990). Evaluating Text-to-Speech Systems: Some Methodological Aspects. *Speech Communication*, 9(4): 263–270.

van Bezooijen, R. and Pols, L. C. W. (1991a). Evaluation of Allophone and Diphone Based Text-to-Speech Conversion at the Paragraph Level. In *Proceedings of the 12th International Congress on Phonetic Sciences*, volume 3, pages 498–501, Aix-en-Provence, France.

van Bezooijen, R. and Pols, L. C. W. (1991b). Performance of Text-to-Speech Conversion for Dutch: A Comparative Evaluation of Allophone and Diphone Based Synthesis at the Level of the Segment, the Word, and the Paragraph. In *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech)*, volume 2, pages 871–874, Genova, Italy.

van Bezooijen, R. and Pols, L. C. W. (1993). Evaluation of Text-to-Speech Conversion for Dutch. In van Heuven, V. J. and Pols, L. C. W., editors, *Analysis and Synthesis of Speech. Strategic Research Towards High-Quality Text-to-Speech Generation*, pages 339–360, Mouton de Gruyter, Berlin, Germany.

van Santen, J. P. H. (1992). Diagnostic Perceptual Experiments for Text-to-Speech System Evaluation. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, volume 1, pages 555–558, Banff, Canada.

van Santen, J. P. H. (1993). Perceptual Experiments for Diagnostic Testing of Text-to-Speech Systems. *Computer Speech and Language*, 7(1):49–100.

van Son, N. and Pols, L. C. W. (1989a). Final Evaluation of Three Multipulse LPC Coders: CVC Intelligibility, Quality Assessment and Speaker Identification. ESPRIT-SPIN Deliverable, IZF Report 17, IFA Report No. 103, 68 pages.

van Son, N. and Pols, L. C. W. (1989b). First Quality Evaluation of a Diphone-Based Synthesis System for Italian. ESPRIT-SPIN Deliverable, IZF Report 15, IFA Report No. 105, 47 pages.

van Son, N. and Pols, L. C. W. (1989c). Intelligibility of Words in Isolation and Words in Semantically Unpredictable Sentences. A Study of Two Diphone-Based Speech Synthesis Systems in French. ESPRIT-SPIN Deliverable, 20 pages + app.

van Son, N. and Pols, L. C. W. (1989d). Review of Synthesis Evaluation Activities during Five ESPRIT/SPIN Years. ESPRIT-SPIN Deliverable, 25 pages.

van Son, N. and Pols, L. C. W. (1989e). Second Quality Evaluation of a French Diphone-Based Synthesis System: Identification and Quality Ratings of Consonant Clusters. ESPRIT-SPIN Deliverable, IZF Report 16, IFA Report No. 104, 51 pages.

van Son, N., Pols, L. C. W., Sandri, S., and Salza, P. L. (1988). First Quality Evaluation of a Diphone-Based Speech Synthesis System for Italian. In *Proceedings of Speech'88, Seventh FASE Symposium*, volume 2, pages 429–436, Edinburgh, UK.

Voiers, W. D. (1977). Diagnostic Evaluation of Speech Intelligibility. In Hawley, M., editor, *Benchmark Papers in Acoustics, Speech Intelligibility and Speaker Recognition*, volume 11, pages 374–387, Dowden, Hutinson, & Ross, Stroudsburg, USA.

Voiers, W. D. (1982). Measurement of Intrinsic Deficiency in Transmitted Speech: The Diagnostic Discrimination Test. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1004–1007, Paris, France.

Voiers, W. D. (1983). Evaluating Processed Speech Using the Diagnostic Rhyme Test. *Speech Technology*, 1(4):30–39.

Vonwiller, J., King, R., Stevens, K., and Latimer, C. (1990). Comprehension of Prosody in Synthesised Speech. In *Proceedings of the Third Australian International Conference on Speech Science and Technology (SST)*, pages 244–249, Melbourne, Australia.

Wery, B. W. M. and Steeneken, H. J. M. (1993). Intelligibility Evaluation of 4-5 KBPS CELP and MBE Vocoders: The Hermes Program Experiment. In *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech)*, pages 67–70, Berlin, Germany.

Zhou, K. C. (1986). Preliminary Evaluation of a Chinese Text-to-Speech System. In *Proceedings of the First Australian International Conference on Speech Science and Technology (SST)*, pages 162–167, Canberra, Australia.

This list is, of course, not complete, but it supplies some relevant keywords for a further search and presents a snapshot of the types of work that have been carried out in the past. For a more complete and up-to-date listing of relevant papers, please consult (and contribute to) the web pages of the ISCA Speech Synthesis Special Interest Group.