

Evaluation of the MU-TALK Speech Synthesis System

Andrew Lampert

December 2004

1	Introduction.....	3
2	An Overview of Speech Synthesis Evaluation	3
3	Taxonomy of Evaluation Techniques	4
4	Further Issues in Text-to-Speech Evaluation.....	7
5	The MU-Talk Text-to-Speech System.....	8
6	Evaluating the MU-Talk Text-to-Speech System.....	10
6.1	Comments on Choosing Human Evaluation Subjects	10
7	Overall MU-Talk TTS System Evaluation	10
7.1	Comprehension Testing	11
7.1.1	Comprehension of Connected Text	11
7.1.2	Comprehension of Isolated Sentences	12
7.2	Speech Quality Evaluation.....	13
7.2.1	Mean Opinion Score	13
7.2.2	Categorical Estimation.....	14
7.2.3	Pair Comparison.....	15
8	Segmental Evaluation	16
8.1	The Diagnostic Rhyme Test.....	16
8.2	Modified Rhyme Test	17
8.3	Phonetically Balanced Word Lists.....	19
8.4	Minimal Pairs Intelligibility Testing.....	21
9	Prosody Evaluation	22
10	Linguistic Evaluation	24
10.1	Text Pre-processing Evaluation	24
10.2	Grapheme to Phoneme Evaluation.....	25
10.3	Context-Sensitive Rule Evaluation.....	26
11	Field Tests	27
12	Conclusion	27
13	References.....	27

1 Introduction

The goal of this work is to examine current methods for speech synthesis evaluation and design an evaluation programme for the MU-Talk Text-to-speech system [17].

In addressing this goal, I first present an overview of the evaluation of speech synthesis and text to speech systems, covering the various different evaluation methodologies that can be applied. In order to structure this exploration of the evaluation space, I explore one of several possible evaluation taxonomies.

Following the general overview of speech synthesis evaluation, I present a brief overview of some common issues that arise in evaluation.

The next section provides an overview of the MU-Talk system, followed by a detailed description of the series of evaluation that are proposed for the evaluation programme.

I finish with some concluding remarks.

2 An Overview of Speech Synthesis Evaluation

In general, humans are conditioned by a lifetime of speaking and listening to speech. We are very sensitive to small changes in speech quality. When listening to speech, we infer much more information than is conveyed through the words alone. From a spoken utterance, a listener can often detect information about the speaker's emotional state, as well as their age, their gender, their accent, and many other characteristics.

The performance of the state-of-the-art in speech synthesis systems remains well below that of human speech in many aspects. It is, of course, debatable whether synthesized speech should ever aim to become indistinguishable from human speech. In general though, this means that listeners must often expend more effort to understand and comprehend synthesized speech than human speech. Especially for users unaccustomed to synthesized speech, listening to a speech synthesizer for extended periods can be both tiring and unsatisfactory.

In order to improve speech synthesis systems and to be able to reliably compare different systems, we must be able to rigorously evaluate and compare speech synthesis systems. Unfortunately, evaluating speech synthesis and text-to-speech (TTS) systems is a difficult task. A single word in text can have many correct pronunciations, so it is much harder to automatically test a synthesizer's phoneme accuracy. Quite apart from this, much of the "quality" metric of a speech synthesis system is based not only on whether pronunciations are correct but also whether it "sounds good". One of the main system-level synthesis evaluation techniques is to have humans listen to the result and respond to specific questions or make subjective judgements.

3 Taxonomy of Evaluation Techniques

Part of the difficulty of evaluating speech synthesis systems also stems from the fact that the evaluation is an inherently multidimensional problem. There are many different types of evaluations that can be made, each with different evaluation goals and requirements. Several attempts have been made to create a standard taxonomy for the types of speech synthesis assessment (see e.g. [5], [11], [25]). The Expert Advisory Group on Language Engineering Standards (EAGLES) Spoken Language Working Group has used these prior attempts to determine a number of useful distinguishing parameters for different assessment techniques. Their proposed assessment taxonomy is shown below in Figure 1.

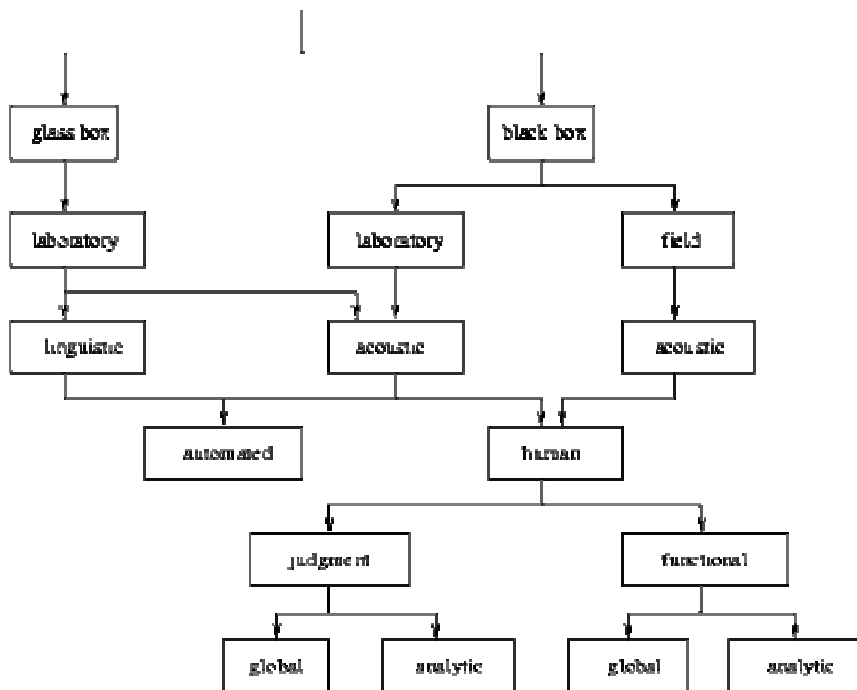


Figure 1: Relationships among dimensions of a taxonomy for speech synthesis evaluation. From Gibbon, Moore and Winski [4]

Although, as mentioned, there are many other possible ways of categorising the various assessment tasks and techniques, I will use the EAGLES assessment taxonomy as a navigation aid for exploring some of the many dimensions of speech synthesis evaluation.

The first categorisation of assessment methods in the chosen taxonomy is between black box and glass box approaches. Black box assessment methods focus on overall system or module functionality, without regard for the internal structure. Glass box assessment methods, on the other hand, generally require knowledge of the internals of a speech system, and focus on testing specific components of a system rather than the system as a

whole. In the case of a speech synthesis system, glass box assessment could focus on a specific module (e.g. the text pre-processing module) or even specific parts of a module, such as how well the text pre-processing module handles homographs.

The distinction between black box and glass box (or white box) testing is one that has been considered extensively in the area of software testing within Software Engineering literature (e.g. [9]). Applying this distinction to speech synthesis evaluation seems natural, given that speech synthesis systems are implemented as modular software systems.

It should be immediately apparent that glass box assessment is likely to be very useful to researchers and developers, but not of much interest to end users of speech synthesis systems. Black box assessment, on the other hand, can be useful for users, researchers and developers. Black box assessment can, for example, allow different systems to be compared, or trace the improvements over time in a single speech synthesis system. Black box assessment will not, however, make it possible to locate the exact modules that are limiting overall system performance. For such diagnostic purposes, glass box evaluations are often required. For a glass box evaluation, either a single module is tested in isolation, or all modules except one are kept constant. Characteristics of the chosen module are systematically varied, while variations in output are observed and recorded. This allows the differences in the synthesis system output to be attributed to specific variations in that module.

The next level of assessment taxonomy distinction is between laboratory and field assessments. As mentioned previously, glass box assessment is only useful in the laboratory, while black box testing can serve both users and developers, and so can be applied in the laboratory and in the field.

The fundamental distinction between laboratory and field tests comes down to the conditions under which evaluations are performed. Laboratory tests are generally performed wherever the system is being developed, though hopefully under controlled conditions. Laboratory tests may be application specific, or may be more general in their focus.

Field tests are always performed in the real environment of use for a specific application. For a speech synthesis system, field tests can only be meaningfully performed when a specific application has been developed around the system, and where users are intended to perform known tasks with the system. The results of field tests cannot, as a rule, be generalised to other systems, applications or environments.

In terms of the three types of evaluation distinguished by Hirschman and Thompson [6], field testing corresponds with Adequacy Evaluation, where the goal is to find out how well a system, or a component of a system, does what it is required to do, relative to the tasks and users at hand. Laboratory testing, on the other hand, tends to correlate with both Diagnostic Evaluation and Performance Evaluation in the Hirschman and Thompson taxonomy.

Returning to the Gibbon, Moore and Winski taxonomy, further distinctions are drawn between evaluation methods that are used to assess the linguistic and acoustic modules within a TTS or speech synthesis system. These evaluations can be either fully automated or human-based assessments.

Particularly at a black box level, human evaluations tend to be more relevant, especially for acoustic module tests or entire system tests. This should not be surprising, given that the value of a speech synthesis system generally comes down to its ability to help end users (humans) perform their intended tasks. It follows logically that those users are best able to assess the system's success or failure to do so.

Unfortunately, human tests are time-consuming and expensive to perform, and usually require a large number of repetitions in order to achieve statistically significant results. For these reasons, there has been significant work in trying to automatically assess speech output. Automatic assessment can include, for example, testing symbolic phonemic transcription component output against a pronunciation dictionary. In such cases, there may be no requirement for a human to assess speech output directly. Other attempts have included measuring acoustic discrepancy between the output of speech synthesis systems and the real human speech that serves as a "gold standard" of what the system is intending to imitate. Automated assessment such as this relies on a clear understanding of how human listeners evaluate differences between two realisations of the same linguistic message. Such a clear understanding of human speech processing has not yet been achieved. Ultimately, there are unfortunately not many components or aspects of a speech synthesis system for which assessment can be easily or reliably automated.

Where human testing is used, performance can be measured objectively or subjectively. Functional (objective) assessment, can measure how well the system achieves its communicative goal, and is usually performed by indirectly inferring communicative success based on the ability of listeners to correctly identify the sounds they have heard. Examples of functional tests include speech intelligibility and comprehension tests.

Judgement testing is by nature subjective. Essentially, it is opinion testing and for speech synthesis evaluation, often involves a group of humans being asked to judge the performance of a TTS system by explicitly rating specific attributes and characteristics of the system on a set of rating scales. As an example, measures of speech naturalness tend to be judgement based. Naturalness is an indication of the extent to which a speech synthesizer sounds like a human - a characteristic that is desirable for most, but not all, applications. Though it might seem counter-intuitive, it is possible to have an artificial-sounding voice that is highly understandable. Similarly, it is possible to have a voice that sounds natural but is difficult to understand (although this is likely to be less common).

Finally, a distinction can be made between global and analytic testing. Judgement tests often include rating scales that cover *global* aspects of system performance including "overall quality", "naturalness" and "acceptability". Functional global testing could include assessing whether the information exchange or task completion for the speech synthesis system is acceptable.

4 Further Issues in Text-to-Speech Evaluation

As already discussed, there are a myriad of issues that make evaluating speech synthesis and text to speech systems a difficult task. Further examples include the need to differentiate assessments based on the nature of the test data used. Van Santen, for example, identified a distinction between the textual test data used in assessing speech synthesis systems [29]. Where fixed lists of words are used to provide test input, the process is referred to as lexically closed, while the use of a text generation algorithm to create textual test input is referred to as lexically open testing. The distinction is important to ensure fairness and objectivity in making comparisons between systems, as requiring comparative tests to use lexically open test data prevents developers from optimising their systems in an ad hoc manner to perform well on specific tests. The point is a valid one – it is difficult to generalize with any degree of confidence from performance on a specific set of carefully selected input data to performance in general.

The same distinction between open and closed data can also be applied to user responses. While closed responses (multiple choice answers) make automated scoring of tests very simple, open responses are usually preferable. One reason for this is that open responses ensure that the intelligibility or comprehension score of random guessing is very low. This is not necessarily true for multiple choice responses.

As discussed later in section 6.1, another well documented consideration is that the experience of listeners in listening to synthetic speech can also greatly affect the results of various types of speech evaluation. Similarly, the repeated presentation of material can also induce a learning effect for listeners. Even without repetition, listeners can make use of a variety of syntactic, semantic, pragmatic, and structural information to help them recover the intended message, even if the synthesized speech itself is not intelligible. These and other effects must be accounted for in the design of evaluation programmes.

For actual speech applications, field tests can cover the range of data that is likely to be encountered for the intended tasks, and are far more important than any standard battery of generic laboratory tests. This is especially true, because it has been shown that the performance of a human listener in recognising and responding to synthesized speech can vary greatly, depending on the complexity of the task being performed [20]. Indeed task complexity can significantly affect intelligibility results for TTS systems. As a result, evaluations involving small, lexically closed sets of input data cannot reliably be directly compared with results from evaluations with more complex, lexically open sets of input data.

5 The MU-Talk Text-to-Speech System

Having discussed the very broad range of possible evaluations that can be carried out for speech synthesis and text-to-speech systems, and some of the general issues to be aware of in evaluation, I now focus on describing the MU-Talk text-to-speech system, which will be the focus for this evaluation. The MU-Talk system (where the “MU” stands for Macquarie University) is the result of several decades of work in the Speech Hearing and Language Research Centre at Macquarie University.

The MU-Talk system has evolved gradually, starting from work performed by John Clark in the mid 1970s on a SID serial formant synthesiser and its accompanying synthesis-by-rule based text-to-speech system. In the mid 1980s, work by Clark, Clive Summerfield and Robert Mannell substituted the SID synthesiser with a parallel formant synthesiser. This work was followed by Clark and Mannell’s development of an accompanying synthesis by rule based TTS system targeting the parallel formant synthesiser. In the mid 1990s additional work was done to experiment with two separate approaches to concatenation synthesis. One approach used formant parameter diphones and the other used channel vocoder parameter diphones. Note that to date no work has been carried out on the much more common LPC and waveform concatenation approaches.

An architecture diagram of the current MU-Talk system is shown in Figure 2. From this diagram, we can see that the main components in MU-Talk are:

- Text Preprocessor Module
- Grapheme to Phoneme Conversion Module
- Context Sensitive Allophonic Analysis Module
- High Level Prosodic Analysis Module
- Low Level Prosodic Analysis Module
- Intrinsic F0 Analysis Module
- Channel Diphone Selection and Concatenation Module
- Formant Diphone Selection and Concatenation Module
- Synthesis By Rule Module
- Channel Synthesis Module
- Formant Synthesis Module

The MU-Talk system is used primarily as a research tool, rather than having been applied in any particular application domain.

MU-Talk Text-to-Speech System

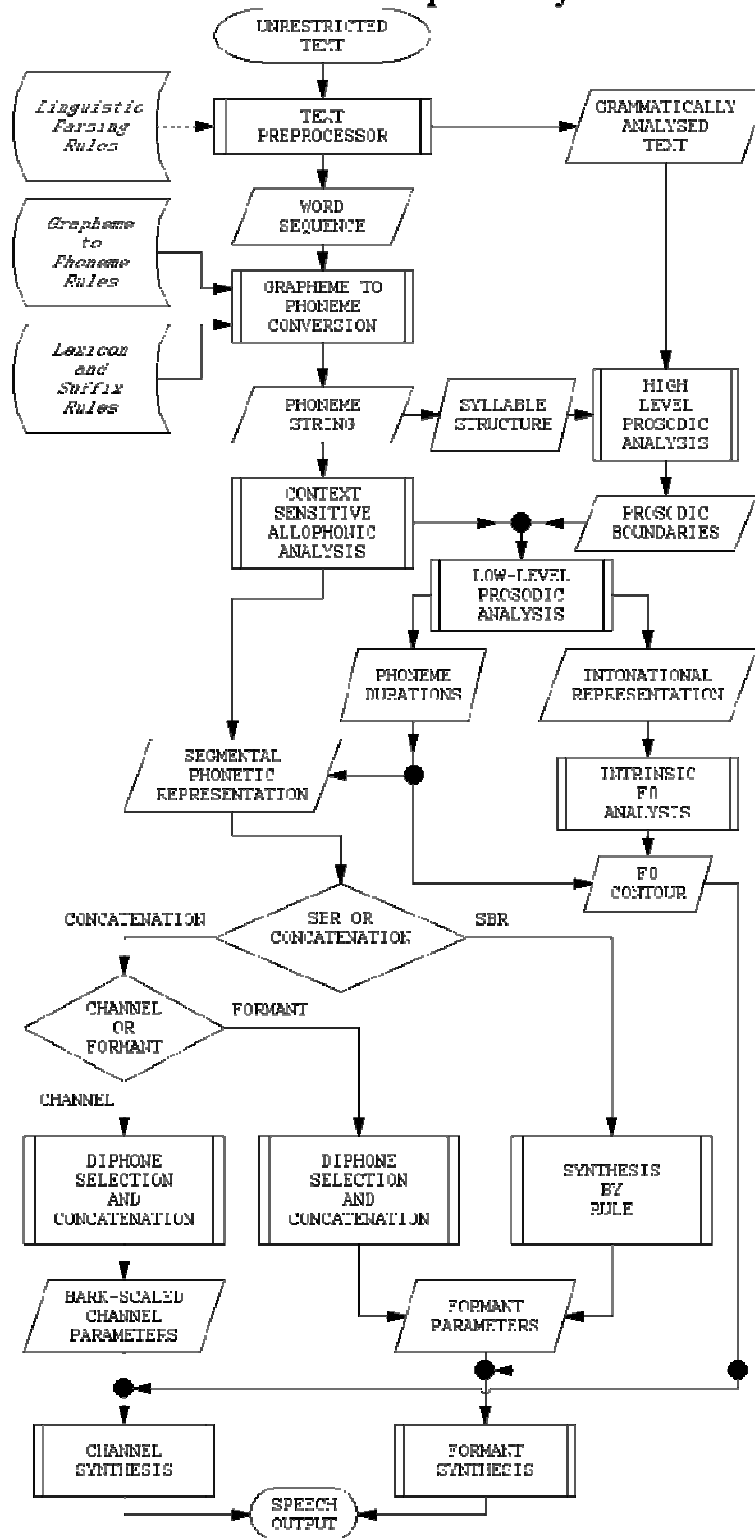


Figure 2: The architecture of the MU-Talk Text-to-Speech System

6 Evaluating the MU-Talk Text-to-Speech System

In the following sections of this document, I present a comprehensive programme of evaluations that could be carried out to assess the performance of the MU-Talk speech synthesis system.

As discussed in section 3, there are many dimensions we could choose to evaluate in the MU-Talk System.

Firstly, I will focus on evaluating the overall MU-Talk system, without regard for specific modules. Finally, I will look at testing individual modules and components.

6.1 *Comments on Choosing Human Evaluation Subjects*

For all tests involving human listeners, it is important to take into account the level of non-expert experience with synthesized speech that each subject has. Various studies (e.g. [7], [8], [12]) have shown that the intelligibility of synthesized speech increases as a result of experience with synthesized speech, even after only a few minutes of exposure. Experience with listening analytically to speech can also boost intelligibility performance [25]. There are indications, however, that the effect of learning depends on the type of synthesis used [12]. This may be significant in testing the MU-Talk system, as it provides modules for both Synthesis By Rule (SBR) and Formant Diphone Concatenation.

In general, subjects for any MU-Talk evaluations should not have a hearing impairment, and should be native speakers of the language being tested.

7 Overall MU-Talk TTS System Evaluation

Overall testing of the MU-Talk system will be performed as black box evaluation, and will focus on testing the performance MU-Talk as a whole system, without considering the performance of individual modules within the system. Because there is no single test to evaluate the adequacy of speech output from a speech synthesis system, a number of complimentary assessments are proposed, each aiming to address a particular aspect of the final speech output.

The types of evaluation that are proposed are comprehension and speech quality assessment.

These and all other listening tests will be conducted under controlled conditions, including:

- Placing subjects in a sound proof room (e.g., Audiology Clinic Rooms at Macquarie University)
- Using standard equipment across all experiments.

7.1 Comprehension Testing

Comprehension tests should be performed with both connected speech and isolated sentences.

7.1.1 Comprehension of Connected Text

Because we don't have a particular target application in mind, the paragraph texts should be relatively broad and varied in genre and style, to provide results that are generic as possible. Such texts could either be extracted from text corpora, or generated using an appropriately configured Natural Language Generation (NLG) system.

As a control measure, human produced speech for each text should also be presented to the user, to provide a reference point for comprehension. For testing the Synthesis By Rule (SBR) aspect of MU-Talk, the human speech should be from the individual on whose speech the table values and synthesis rules are based. Equally, for MU-Talks concatenative synthesis, the relevant person would be the individual from whose voice the basic concatenation components are constructed.

Method

The comprehension test with connected text should be setup and run as follows:

- Paragraph sized texts are presented to the listener using synthesized speech.
- A series of questions, related to the text content, are posed.
- The user responds to each question.

The questions should be open response questions, rather than closed response questions. Open response tests are known to be more sensitive, and are therefore more likely to highlight differences between test conditions. Care should be taken to ensure that answers are unambiguously right or wrong.

As well as the user's response to each question, response times should also be recorded. Response time is often higher for synthesized speech than human speech. Studies have shown that this seems to be because humans make more of an effort to understand synthesized speech.

Ideally, different groups of subjects should be used, so that no listener hears more than one version of the same text. Additionally, each listener should listen to an equal number of different text versions.

When assessing the results of this test, the comprehension results should be compared with the results for human speech, to abstract away from the intrinsic difficulty of the questions and topic.

7.1.2 Comprehension of Isolated Sentences

For comprehension of isolated sentences, I propose that a Sentence Verification Test be used. Such approaches have been used for many comprehension studies (e.g. [20], [22]). For each of the sentence comprehension tests, no more than 7 words per sentence are used, in order to reduce cognitive load on listeners.

Method

In sentence verification tests, subjects are required to respond ‘true’ or ‘false’ after listening to each sentence. For the MU-Talk tests, sentences presented should have 3 words each, and be true/false statements. The sentences should be pre-tested to determine whether the final word in each sentence is predictable. Sentences with both low-predictability and high-predictability should be used, and the results between these differing sentence types compared.

For each presented sentence, the response time between presentation and the user response is recorded, along with whether the user response was correct or incorrect. Results would be expected to show patterns between response time and the degree of predictability of the sentences. Additional experimentation could involve the use of sentences with different numbers of words.

As a control measure, the same sentences should be recorded and played back with a natural human voice. As already discussed, this human voice should be the same person’s voice that is used to drive the synthesis system.

Either or both of the Harvard Psychoacoustic Sentences or Haskin Sentences could be used to assess comprehension at the sentence level.

The Harvard sentences are chosen so that the various segmental phonemes of English are represented in the frequency they occur in natural speech. The sentences are all meaningful, and contain a wide range of syntactic constructions. As a closed set of 100 sentences, however, there are potential issues with the learning effect for listeners, as has been documented in [18]. The first five sentences of the Harvard corpus are:

- The birch canoe slid on the smooth planks.
- Glue the sheet to the dark blue background.
- It’s easy to tell the depth of a well.
- These days a chicken leg is a rare dish.
- Rice is often served in round bowls.

Because the Harvard sentences are semantically meaningful, listeners can perform well on comprehension tests, even if they don't understand parts of the sentence. Complimentary to the Harvard sentences are the Haskin sentences, in which the test sentences are not meaningful. As a result, unlike the Harvard sentences, missed items cannot be concluded from the context as easily, if at all. Like the Harvard sentences, the Haskin sentences are a closed set, meaning that again test listeners can be presented with each sentence only once, to ensure reliable results that are not influenced by learning effects. The natural extension to Haskin sentences are Semantically Unpredictable Sentences (SUS) [1], which are generated on the fly, based on several sentence templates and a corpus of words with known parts-of-speech. This can, to some degree, overcome learning problems from using a closed set of sentences.

7.2 Speech Quality Evaluation

One of the main measures of overall system performance is measured by listeners' subjective impression of speech quality. In most cases, speech quality is a very broad term that encompasses everything for how natural the speech sounds, to how intelligible and comprehensible it is. It must be noted that assessments of overall speech quality are highly dependent on the particular texts used. Also, perception of speech quality is influenced by both linguistic and psychological factors, including factors such as the genre of text used and the presentation sequence.

Perhaps particularly interesting is that some studies [23] have shown that there can be little to no correlation between speech intelligibility and acceptability ratings for synthesized speech. Other studies have shown the opposite. Ultimately, perhaps all we can say is that subjective acceptability is not necessarily a simple consequence of intelligibility, and that a distinction needs to be made between the aesthetic and functional aspects of synthetic speech. Speech quality evaluation is generally seen to be the best measure of the aesthetic aspects. In evaluating speech quality, there are numerous available methodologies to choose from.

7.2.1 Mean Opinion Score

The Mean Opinion Score (MOS) is probably the most widely used and simplest method to evaluate the overall quality of a speech synthesis system. Because of its simplicity and ubiquity, it seems a worthwhile part of the evaluation of the MU-Talk system.

Method

The MOS methodology generally includes a 5 level scale for ranking system performance, and the user's task is simply to evaluate the synthesized speech they have heard on this scale. The 5 levels could be:

1. Excellent
2. Good

3. Fair
4. Poor
5. Bad

In general the texts used for performing the MOS test should be of the same genre and length as the target texts that will be synthesized by the system in the field. In the case of MU-Talk, where we don't have a particular task context in mind, the text presented to the user should be of varying lengths, and genre. Comparisons of different combinations of text length and genre could produce interesting quality metrics of the MU-Talk system, and also should allow wider comparison with other systems, on the basis of them having been evaluated with text of the same genre and length.

7.2.2 Categorical Estimation

In categorical estimation, rather than just assessing the overall speech quality, listeners are asked to quantify various attributes or aspects of the speech separately. Possible attributes for the Categorical Rating Test (CRT) include:

- Overall Impression
- Intelligibility
- Comprehensibility
- Acceptability
- Listening Effort Required
- Naturalness
- Friendliness
- Politeness
- Liveliness
- Pronunciation
- Speed
- Distinctness
- Stress
- Voice Pleasantness

Note that some of these attributes relate more to specific aspects of speech output than to overall quality, and as a result require analytic listening. All these judgements about speech output, however, can be useful in determining the overall perception of speech quality for each listener.

Method

As for the Mean Opinion Score tests, listeners are presented with varying length texts from a variety of genres, and asked to assess the speech in terms of the characteristics noted. The assessments are made by rating the speech on a scale, ranging from poor to excellent for each characteristic. A 10 point scale should be used to ensure the scale is

wide enough to allow differentiation amongst different systems. The scale should have meaningful explanatory labels for each attribute. In the case of MU-Talk, our control system is again human recorded speech. If time and resources permitted, both the MOS and Categorical Estimation tests could be performed with the Festival Speech Synthesis system, for purposes of system comparison.

7.2.3 Pair Comparison

Pair comparisons involve side-by-side testing of two different speech synthesis systems. For MU-Talk, an obvious pair comparison could be performed with the Festival speech synthesis system, which is freely available. Pair comparisons provide an easy, if subjective way for comparing overall performance of different systems.

Method

The same textual test data is used for both systems. Listeners are presented with the same data in synthesized speech from each system sequentially. Repeated presentation of the same text should produce a learning affect, which should thus result in higher intelligibility, and possibly a more positive attitude towards those later systems. In order to control confounding from order effects, the order in which the two systems' output is presented must be counter balanced according to a Graeco-Latin square across listeners. This ensures that all possible orders of versions were presented equally often across listeners.

Listeners are required to indicate their preferred presentation from each pair. The category "equal" is not allowed. If more than two systems are being compared, as will be in the MU-Talk evaluation programme where MU-Talk output, Festival output and human speech is presented, then the same spoken stimuli from each synthesizer is presented in random order, for each possible pair combination across each sentence or text passage.

8 Segmental Evaluation

Having evaluated the overall quality and performance of the MU-Talk system, the evaluation focus should then move onto assessing more detailed aspects of particular modules in the system. The first series of these drill-down evaluations are related to speech intelligibility, which of course also plays a significant role in overall system quality.

With segmental evaluation methods, single segments or phonemes are tested separately for intelligibility. The most commonly used methods to test intelligibility of synthetic speech involve the use of rhyme tests and nonsense words. The rhyme tests have several advantages [10]. Firstly, the number of stimuli is reduced and the test procedure can be performed in a conveniently short space of time. Also of particular importance is that non-expert listeners can participate without having to be trained, and reliable results can be obtained with relatively small subject groups (e.g., from 10 or 20 subjects). Rhyme tests are therefore relatively easy and economic to conduct.

In general, the intelligibility metric is simply the number of correctly identified words as a proportion of all words. There are many aspects of a speech synthesis system that can act to reduce intelligibility, including the computation and production of incorrect or inappropriate suprasegmental information, incorrect grapheme to phoneme conversion and errors in converting allophones to waveforms. To help localise the source of intelligibility problems, rhyme tests can provide diagnostic information through confusion matrices. Confusion matrices provide information about how different phonemes are misidentified and help to identify the source of specific problems for developers.

Despite their appeal and almost universal acceptance, rhyme tests also have some significant disadvantages. With rhyme tests that use lists of monosyllabic words, only single consonants are tested. Usually, the vocabulary is also fixed and public so the system designers may tune their systems to perform well on the specific rhyme test data, and listeners might reasonably remember the correct answers if they participate in the test more than once.

8.1 *The Diagnostic Rhyme Test*

As already mentioned, the most common method for assessing intelligibility of synthesized speech is to use rhyme tests. Probably the most well known rhyme test is the Diagnostic Rhyme Test or DRT [29, 30], based on work by Fairbanks [1]. The DRT uses a set of 96 word pairs, which differ by a single acoustic feature in the initial consonant. From the set of acoustic features shown in Table 1, one feature is chosen to differ between each word in a pair. Only one of the words from each pair is randomly selected and presented. The listener is asked to identify which word from each pair they have heard.

Table 1: Features used to distinguish between each word in a pair

Feature	Example Word Pairs
Voicing	veal – feel, dense – tense
Nasality	reed – deed
Sustension	vee – bee, sheet – cheat
Sibiliation	sing – thing
Graveness	weed – reed
Compactness	key – tea, show – sow

DRT can provide valuable information about diagnosing problems with production of initial consonants. It does not, however, provide any evaluation of vowels or prosodic features, and so is only one aspect of testing overall system intelligibility.

Rather than use the DRT for MU-Talk, however, I propose that the Modified Rhyme Test, which is essentially an enhanced DRT, be used instead.

8.2 Modified Rhyme Test

The Modified Rhyme Test (MRT) extends the Diagnostic Rhyme Test by testing both initial and final consonant intelligibility. The test consists of 50 sets of 6 single-syllable words. From each set of 6 words, one word is randomly selected and played, and the user must again decide which word they have heard.

Each word is constructed from a consonant-vowel-consonant sound sequence, and the six words in each set differ only in the initial or final consonant sound. The standard set of words for the MRT is shown in Table 2.

Method

As for all statistical intelligibility tests, a carrier sentence should be used to present test words (for example, “Please write <test word> now”). The test word is spoken without emphasis, and the sentence is the same for each test word. The carrier sentence assures that the reverberant field is excited prior to the test word being spoken, so that its effects are properly accounted for in the evaluation. It also allows dynamics processors such as automatic gain controls or compressors to activate and stabilize.

Each listener’s response can be elicited in one of two ways: either the user can choose the word they have heard from a closed set of words, or the user can write the word they have heard. The first method is the easiest to score automatically, and is the method usually adopted for MRT. Because of its ease of assessment, I would choose to use a closed response mechanism for this evaluation.

In order to ensure maximal listener uncertainty, listeners would only be presented with the set of alternative answers a few seconds after auditory presentation. This means that

when the listener hears the carrier sentence, they do not know whether the critical phoneme is word-initial, word-internal or word-final. This makes the evaluation more reflective of natural listening situations.

Listener responses are scored on:

- The number of words heard correctly/incorrectly; and
- The response time for each user response.

For synthesis evaluation, the frequency of particular consonant confusions is perhaps the most useful diagnostic information. The percentage of words correctly heard makes a reasonable benchmark of system intelligibility performance.

Table 2: The set of 300 Stimulus Words for the MRT

went	sent	bent	dent	tent	rent
hold	cold	told	fold	sold	gold
pat	pad	pan	path	pack	pass
lane	lay	late	lake	lace	lame
kit	bit	fit	hit	wit	sit
must	bust	gust	rust	dust	just
teak	team	teal	teach	tear	tease
din	dill	dim	dig	dip	did
bed	led	fed	red	wed	shed
pin	sin	tin	fin	din	win
dug	dung	duck	dud	dub	dun
sum	sun	sung	sup	sub	sud
seep	seen	seethe	seek	seem	seed
not	tot	got	pot	hot	lot
vest	test	rest	best	west	nest
pig	pill	pin	pip	pit	pick
back	bath	bad	bass	bat	ban
way	may	say	pay	day	gay
pig	big	dig	wig	rig	fig
pale	pace	page	pane	pay	pave
cane	case	cape	cake	came	cave
shop	mop	cop	top	hop	pop
coil	oil	soil	toil	boil	foil
tan	tang	tap	tack	tam	tab
fit	fib	fizz	fill	fig	fin
same	name	game	tame	came	fame
peel	reel	feel	eel	keel	heel
hark	dark	mark	bark	park	lark
heave	hear	heat	heal	heap	heath
cup	cut	cud	cuff	cuss	cud
thaw	law	raw	paw	jaw	saw
pen	hen	men	then	den	ten

puff	puck	pub	pus	pup	pun
bean	beach	beat	beak	bead	beam
heat	neat	feat	seat	meat	beat
dip	sip	hip	tip	lip	rip
kill	kin	kit	kick	king	kid
hang	sang	bang	rang	fang	gang
took	cook	look	hook	shook	book
mass	math	map	mat	man	mad
ray	raze	rate	rave	rake	race
save	same	sale	sane	sake	safe
fill	kill	will	hill	till	bill
sill	sick	sip	sing	sit	sin
bale	gale	sale	tale	pale	male
wick	sick	kick	lick	pick	tick
peace	peas	peak	peach	peat	peal
bun	bus	but	bug	buck	buff
sag	sat	sass	sack	sad	sap
fun	sun	bun	gun	run	nun

The MRT has been used for comparing the performance of different synthesizers by several researchers, including Logan et al. [15]. As such, it could be interesting to compare the results from an MRT test with MU-Talk, with the results from other similar text-to-speech systems.

8.3 Phonetically Balanced Word Lists

Using phonetically balanced word lists is an established technique for assessing speech intelligibility that was developed at Harvard University in the 1940s, and such an approach provides another means of testing speech intelligibility with the MU-Talk system. The test involves the presentation of a set of twenty phonetically balanced, monosyllabic word lists, the first four of which are shown in Table 3. These phonetically balanced word lists offer phonemes in the approximate distribution that they appear in normal speech. Such lists have been used very widely in statistical intelligibility testing.

It should be noted that phonetically balanced intelligibility testing requires more training of listeners than other statistical tests, and the technique is apparently very sensitive to signal-to-noise ratios, in that a relatively small change in the signal to noise ratio causes a large change in the intelligibility score.

Method

Words from each phonetically balanced word list are presented in a new, random order each time the list is used, and each word is spoken in a carrier sentence. The relative difficulty of the words is constrained so that items that are always missed or always correct are removed, leaving only those words that provide useful information. An open

response set should be used, meaning that users are required to record what they have heard, rather than choose between a closed set of options.

Table 3: The first four phonetically balanced word lists

are	awe	ache	bath
bad	bait	air	beast
bar	bean	bald	bee
bask	blush	barb	blonde
box	bought	bead	budge
cane	bounce	cape	bus
cleanse	bud	cast	bush
clove	charge	check	cloak
crash	cloud	class	course
creed	corpse	crave	court
death	dab	crime	dodge
deed	earl	deck	dupe
dike	else	dig	earn
dish	fate	dill	eel
end	five	drop	fin
feast	frog	fame	float
fern	gill	far	frown
folk	gloss	fig	hatch
ford	hire	flush	heed
fraud	hit	gnaw	hiss
fuss	hock	hurl	hot
grove	job	jam	how
heap	log	law	kite
hid	moose	leave	merge
hive	mute	lush	lush
hunt	nab	muck	neat
is	need	neck	new
mange	niece	nest	oils
no	nut	oak	or
nook	our	path	peck
not	perk	please	pert
pan	pick	pulse	pinch
pants	pit	rate	pod
pest	quart	rouse	race
pile	rap	shout	rack
plush	rib	sit	rave
rag	scythe	size	raw
rat	shoe	sob	rut
ride	sludge	sped	sage
rise	snuff	stag	scab

rub	start	take	shed
slip	suck	thrash	shin
smile	tan	toil	sketch
strife	tang	trip	slap
such	them	turf	sour
then	trash	vow	starve
there	vamp	wedge	strap
toe	vast	wharf	test
use	ways	who	tick
wheat	wish	why	touch

8.4 Minimal Pairs Intelligibility Testing

In order to further evaluate the MU-Talk system under lexically open conditions, I also propose to use the Minimal Pairs Intelligibility (MPI) test, developed at Bell Labs [26], [27]. This test was developed as a minimal pairs test to maximise the coverage of the phoneme space, and does so by presenting an algorithm for generating sets of minimal word pairs. Because it remains a minimal pairs test, MPI can still be automatically scored in the same way as DRT results, and the results are equally straightforward to interpret. For these and other reasons, it is an attractive evaluation which complements the earlier MRT testing for MU-Talk.

Apart from being lexically open, Minimal Pairs Intelligibility differs from DRT because it generates minimal pairs that differ by both one and two phonetic features, by deletion or by insertion. DRT uses pairs that always differ by one feature. To ensure broad coverage of testing, MPI test data explicitly contains consonants and vowels, onsets, nuclei and/or codas, consonant clusters, mono-syllabic and poly-syllabic words, and stressed and unstressed syllables.

Method

For each listener, all speech versions occur the same number of times with all combinations of word location in the presented sentence, difference type (i.e., consonant substitution, vowel substitution, insertion, deletion, and 1 vs. 2 feature contrasts), and within-word location (i.e., word initial, word final, or within word). The actual words used are taken from Appendix B of [27], which contains sets of word pairs for each combination of part of speech, difference type, within-word difference location and one and two feature differences.

Word pairs are selected from the lists at random, and one word is inserted into a semantically unpredictable carrier sentence (SUS) [1], positioned in the sentence according to its part of speech. The remaining slots in the carrier sentence are filled with unused words from the minimal pairs lists. Thus the presented sentences are syntactically valid, but semantically meaningless.

These sentences are presented to each listener one by one. Half a second before each sentence is presented, the listener is played a warning tone. As for the MRT tests, listeners are only presented with the set of alternative answers a few seconds after auditory presentation. This means that when the listener hears the carrier sentence, they do not know which word position in the sentence is critical, nor whether the critical phoneme is word-initial, word-internal or word-final.

9 Prosody Evaluation

In general, there hasn't been as much work on evaluating prosody for speech synthesis systems. This almost certainly stems from the difficulty of performing such evaluation. Prosody cannot be assessed in terms of automatically computed error rates, since there is no such thing as "correct" prosody. Even with a gold standard for prosody for specific utterances, measurements of deviations from such a reference are not especially useful, given the non-linear relationship between acoustic and auditory components of prosody [24]. There is, however, a need to be able to assess prosody, as it plays a particularly important role in determining the naturalness of speech output. Thankfully, even though there is no prosodic "gold standard" for any given text (as two completely different realisations can be equally acceptable), listeners are usually able to distinguish consistently between good and bad prosody, which indicates that, at least at some level, the quality of prosody can be measured [24].

In MU-TALK, utterance prosody properties are classified into four primary categories, based on intonational models proposed by Pierrehumbert [1], [18]. These categories are: Stress, Phrasing, Tune Choice, and Pitch Control. These prosodic properties are combined to obtain the final fundamental frequency definitions, which are then time-aligned, smoothed, interpolated, and generated.

For the purposes of evaluating prosody in the MU-Talk system, I propose to use the Prosody Unveiling through Restricted Representation (PURR) method, developed by Sonntag and Portele [24]. This allows for evaluation of the prosodic output of a speech synthesis system independently from its segmental quality.

The approach of the PURR method is to de-lexicalise the speech stimuli to ensure that the listener perceives only the prosody of an utterance. This is done by reducing the speech signal to produce stimuli that convey only intensity, F0 contour and temporal structure.

Method

Textual test data should be of varying genres and nature. Ideally, there should be a mix of declarative statements, questions and requests. Each of these texts should be synthesized by MU-Talk, then de-lexicalised. Each text should also be spoken by a human (again, preferably the human on whose voice the synthesis is modelled), and the same de-lexicalisation process applied to the recorded speech.

These pairs of stimuli for each sentence should be presented sequentially. The order in which the two outputs are presented must be counter balanced according to a Graeco-Latin square across listeners. This ensures that all possible orders of versions were presented equally often across listeners.

Listeners are required to indicate their preferred presentation from each pair. The category “equal” is not allowed.

Because, almost certainly, human prosody will be preferred over MU-Talk’s, it would also be highly desirable to add another synthesis engine into the comparison. Festival is again a good candidate.

Following the proposal from Sonntag and Portele [24], I would use a magnitude estimation test procedure to quantify prosody comparisons across multiple synthesis systems. Magnitude estimation procedures are also used in speech quality assessment, and rely on an assumption that each listener can use a reliable internal reference system to intuitively relate relative quantities to each other. This relational ordering can refer to different properties of stimuli (as in the case of assessing speech quality), and is supposed to be more exact the more intuitively it is done [24]. To make the prosody evaluation process as intuitive as possible, listeners should be asked to assign a line of certain length to each stimuli presented. The length of each line should represent the perceived naturalness of the prosody of each utterance.

To assess listeners’ responses, the lines would be measured and normalised.

10 Linguistic Evaluation

In this section of the MU-Talk evaluation programme, we will focus on evaluation methods which can be used for assessing the intermediate output produced by specific linguistic modules in the MU-Talk system. The specific modules considered will include the text pre-processor and the grapheme-to-phoneme conversion module.

10.1 Text Pre-processing Evaluation

Text pre-processing is a complex task that is often performed using hand-crafted heuristic rules. The process is necessarily highly language dependent. In the case of MU-Talk, which to date has targeted only English, the multi-lingual aspects of text pre-processing will not be assessed.

Examples of the types of tasks that are performed during text pre-processing include the expansion of the following textual features into full words:

- Digits (e.g., Phone numbers);
- Numbers (221);
- Fractions ($2/3$, $1/4$)
- Ordinal Numbers (1st, 2nd, 3rd, 4th)
- Roman Numerals (I, MCMXCVIII)
- Dates (1998, 2004);
- Abbreviations (CSIRO, NATO);
- Acronyms (e.g., i.e., etc.)
- Special Symbols (#, \$21.43, 25%)

Of course, such expansions are often ambiguous, or require contextual knowledge to correctly interpret and expand the text into full words.

There are, unfortunately, no standardised tests for determining the adequacy of text pre-processors. Given that we have identified a number of classes of orthographic transformations, however, it seems possible to use these classes as categories for quantifying text pre-processing performance. Laver et al. [13, 14] took a similar approach in their assessment of the CSTR text pre-processor, by quantifying the errors in various categories.

Method

In general, evaluating the text pre-processor requires a corpus of known inputs and expected outputs. This allows automated testing to be performed across a range of input categories. There is not yet an agreed taxonomy for distinguishing error classes for text pre-processing, but such categories could include each of task types identified in the list

above (dates, abbreviations, acronyms etc.). For each class of input, the number of errors in output would be expressed as a percentage of total inputs. Errors would be determined automatically through comparison of the module output with the gold standard corpus.

There are several aspects of text pre-processing that MU-Talk does not yet handle. Naturally, these would not be evaluated. The following aspects of text pre-processing are not currently implemented in MU-Talk, and would be excluded from evaluation:

- Homographs – Currently, there is no facility in MU-Talk to make choices between alternate pronunciations for homographs. Instead, the first pronunciation is selected by default.
- Parsing/POS Tagging – MU-Talk makes no attempt to parse or identify parts of speech for incoming text.
- Dates – Related to the lack of homograph processing, MU-TALK currently has no way of determining whether "1999" is to be pronounced as a year or an ordinal number.

10.2 Grapheme to Phoneme Evaluation

The grapheme to phoneme module in MU-Talk takes the output of the text pre-processor as input, and processes that text in order to output a representation of the corresponding sequence of sounds. This output takes the form of a series of phonemic symbols.

In order to test the grapheme to phoneme process, firstly a corpus of input words with their correct pronunciation is required. Such a corpus could be based around the Macquarie Dictionary, which includes Australian English pronunciation for words. This is probably the most relevant pronunciation corpus for MU-Talk evaluation.

Method

Words from the Macquarie Dictionary are input to the grapheme to phoneme component, and the output phonemic string is compared with the acceptable pronunciations, which are also extracted from the Macquarie Dictionary. Following the suggestion of Mannell and Clark in [16], the transcription of each word from the grapheme to phoneme module is scored correct, almost correct (only schwa/vowel mismatches) or incorrect. The performance of the grapheme to phoneme module could then be expressed in terms of the percentage of words correctly transformed into a correct phonemic string.

As reported in [16], an additional score that could be computed is a corpus weighted score. To compute such a metric requires the use of a wide coverage English lexicon, such as the Brown corpus. From such a corpus, the relative frequency of each word in English text (of a particular genre) is used to weight the scores for each word. In this way, a weighted average for MU-Talk can be calculated, which places more emphasis on correctly transforming words that occur more frequently in natural language. Such a weighted score potentially gives a more useful guide as to the performance of the grapheme to phoneme module.

Additional diagnostic information could be provided by classifying the types of errors made. Possible types could include: loan word (foreign word) errors, schwa errors, incorrect identification of morpheme boundaries. This set of possible error types is far from exhaustive. Unfortunately, however, there seems to be a lack of well-defined and accepted taxonomies for classifying grapheme to phoneme conversion errors.

10.3 Context-Sensitive Rule Evaluation

The Context Sensitive Rule (CSR) module of MU-Talk has the task of determining the appropriate allophone for each phoneme, as determined by its context. The CSR module examines the context of each phoneme and applies modifications to the non-contrastive features of each phoneme. A non-contrastive feature modification is a one that does not transform a phoneme into another phoneme. Such non-contrastive features might include:

- voicing in approximants;
- different levels of stress in vowels;
- variations in the relative durations of sub-phonemic features (eg. targets, transitions, stop occlusions, stop bursts and aspirations, etc.); and
- non-contrastive shifts in place of articulation (eg. velar to palatal shifts in /k/ and /g/).

Clearly, the performance of this module has a significant impact on the overall naturalness and quality of speech output, so it's contribution to the overall system would also be captured in the subjective speech quality evaluation.

At present, the CSR module is effectively a pre-processor to the synthesis-by-rule branch of MU-Talk. It provides very detailed, and SBR-specific, information regarding the precise quality of each of the allophones.

Method

It seems difficult to assess the contribution of the Context-Sensitive Rule module in isolation. It could be possible, if a carefully constructed corpus of phonemes for various text types (questions, declarations, requests etc.) was constructed, where each phonemic sequence contained an aligned set of acceptable phonetic transcriptions, complete with stress markings, and relevant sub-phonemic features. This would allow the output of the CSR module to be compared with a gold-standard, and its accuracy measured accordingly. Given the difficulty of obtaining agreement on a set of acceptable non-contrastive phoneme features, however, it is doubtful whether such a corpus could actually be constructed.

11 Field Tests

Field tests evaluate the performance of systems within the context of specific users, tasks and environments with which the system will be deployed. In the case of MU-Talk, because we have no specific context of use in mind, field tests are not an applicable or relevant part of the proposed evaluation programme.

12 Conclusion

I have presented a comprehensive and complimentary evaluation programme to assess the current state of the MU-Talk system. The programme has been customised to suit the current architecture and implementation state of MU-Talk, and covers both system level and diagnostic testing.

Each evaluation methodology proposed has been justified and described in terms of how it fits into the overall evaluation framework, and the issues and limitations involved in performing the evaluation.

To execute the entire evaluation programme would obviously be an expensive and time consuming process. Without expert knowledge of the priorities for the MU-Talk developers, I haven't attempted to prioritise the evaluations proposed. Many of the tests can be carried out independently, however, so parts of this evaluation programme which were deemed to be of high importance could be selectively performed, without the need to address all aspects of MU-Talk's evaluation which are described here.

13 References

1. Beckman, M. and Pierrehumbert, J. (1986) "Intonational Structure in Japanese and English" *Phonology Yearbook III*, 15-70.
2. Benoit, C., Grice, M. & Hazan, V. (1996) "The SUS test: a method for the assessment of text to speech synthesis intelligibility using Semantically Unpredictable Sentences". *Speech Communication*, 18, 381-392.
3. Fairbanks, G., (1958). "Test of phonemic differentiation: The Rhyme Test," in *Journal of the Acoustic Society of America*, 30, 596-600.
4. Gibbon, D., Moore, R., Winski, R (eds). (1997). "Handbook of Standards and Resources for Spoken Language Systems", Walter de Gruyter & Co., Berlin.
5. Goldstein, M. (1995). "Classification of methods used for assessment of text-to-speech systems according to the demands placed on the listener", *Speech Communication* Volume 16, 225-244.

6. Hirschman, J., and Thompson, S. (1997). "Overview of Evaluation in Speech and Natural Language Processing", in Cole, R et al (Eds), *Survey of the State of the Art in Human Language Technology*, Cambridge University Press, Cambridge, 409-414.
7. Howard-Jones, P. (1992a). "SOAP, Speech Output Assessment Package". Version 4.0, ESPRIT SAM-UCL-042.
8. Howard-Jones, P. (1992b). "Specification of listener dimensions. : ESPRIT Project 2589 (SAM)", in *Multilingual speech input/output assessment, methodology and standardisation*. University College London, London. Stage report So. 8, Part One, Final report, Year three: 1.III.91-28.II.1992.
9. Howden, W. (1980). "Functional program testing", *IEEE Transactions on Software Engineering*, Volume 6, 162-169.
10. Jekosch U. (1993). "Speech Quality Assessment and Evaluation", In *Proceedings of Eurospeech 93* (2): 1387-1394.
11. Jekosch, U., and Pols, L. (1994). "A feature-profile for application-specific speech synthesis assessment and devaluation", *Proceedings of the 3rd International Conference on Spoken Language Processing*, Yokohama.
12. Jongenburger, W. and Van Bezooijen, R. (1992). "Evaluatie van ELK: Attitudes van de gebruikers, verstaanbaarheid en acceptabiliteit van de spraaksynthese, bruikbaarheid van het zoekstelsel". *Stichting Spraaktechnologie*, Utrecht.
13. Laver, J., McAllister, J., McAllister, M., and Jack, M. (1988). "A Prolog-based automatic text-to-phoneme conversion system for British English.", In *Proceedings of the Second Symposium on Advanced Man-Machine Interface through Spoken Language*, November 19-22, Hawaii.
14. Laver, J., McAllister, M. and McAllister, J. (1989). "Pre-processing of anomalous text-strings in an automatic text-to-speech system.", In Ramsaran, S., *Studies in the pronunciation of English: A commemorative volume in memory of A.C. Gimson*. Croon Helm, London.
15. Logan, J., Pisoni, D., & Greene, B. (1989). "Measuring the Segmental Intelligibility of Synthetic Speech: Results from Eight Text-to-Speech Systems", *Journal of the Acoustic Society of America*, vol 86, 566-581.
16. Mannell R.H., and Clark, J.E., (1987), "Text-to-speech rule and dictionary development", *Speech Communication* 6, North-Holland, pp317-324
17. MU-Talk Text-to-Speech System Website, <http://www.ling.mq.edu.au/home/rmannell/mutalk/main/index.html> , retrieved on 4/12/2004.
18. Pierrehumbert, J. (1981) "Synthesizing Intonation" *Journal of the Acoustical Society of America*, 70, 985-995.
19. Pisoni, D., and Hunnicutt, S., (1980) "Perceptual evaluation of MITalk: the MIT unrestricted text-to-speech system," In *Proceedings of ICASSP-80*, 572-575.
20. Pisoni, D. (1997) "Perception of Synthetic Speech", in Van Santen, J., Sproat, R., Olive, J. and Hirschberg, J.n (Eds), *Progress in Speech Synthesis*, Springer-Verlag, New York, USA, 541-560.
21. Pols, L. (1991). "Quality assessment of text-to-speech synthesis-by-rule", in Furui, S and Sondhi, M., *Advances in speech signal processing*, Marcel Dekker Inc., New York, 387-416.

22. Reynolds, M., Isaacs-Duvall, C., Haddox, M. (2002). "A Comparison of Learning Curves in Natural and Synthesized Speech Comprehension", in *Journal of Speech, Language and Hearing Research*, Vol. 45 (Aug 2002), 802-810.
23. Sluijter, A., Bosgoed, E., Kerkhoff, J., Meier, E., Rietvald, T., Sanderman, A., Swerts, M., Terken, J., "Evaluation of speech synthesis systems for Dutch in telecommunication applications", In *Proceedings of the 3rd ESCA Workshop on Speech Synthesis*, 1998.
24. Sonntag, G. and Portele, T. (1998), "PURR – A method for prosody evaluation and investigation", *Computer Speech and Language*, 12:437-451.
25. Van Bezooijen, R. and Pols, L., (1990). "Evaluating text-to-speech systems: Some methodological aspects" *Speech Communication*. Volume 9, 263-270.
26. Van Santen, J. (1992), "Diagnostic perceptual experiments for text-to-speech system evaluation", In *Proceedings of the Second International Conference on Spoken Language Processing*, Danff, Alberta, Volume 1, p 555-558.
27. Van Santen, J. (1993), "Perceptual experiments for diagnostic testing of text-to-speech systems", *Computer Speech and Language*, 7:49-100.
28. Van Santen, J. (1998) "Evaluation", in Sproat, R (Ed), "*Multilingual Text-To-Speech Synthesis. The Bell Labs Approach*", Kluwer Academic Publishers, Massachusetts, USA.
29. Voiers W. (1977) "Diagnostic Evaluation of Speech Intelligibility." In M.E. Hawley, Ed, *Speech Intelligibility and Speaker Recognition*, Dowden, Huchinson, and Ross; Stroudsburg, PA.
30. Voiers, W., Sharpley, A., and Hehmsoth, C. (1972). "Research on diagnostic evaluation of speech intelligibility" Research Report AFCRL-72-0694, Cambridge Research Laboratories.
31. Yvon, F., Boula de Mareüil, P., d'Alessandro, C., Auberge, V., Bagein, M., Bailly, G., Béchet, F., Foukia, S., Goldman, J-F., Keller, E., O'Shaugnessy, D., Pagel, V., Sannier, F., Véronis, J., and Zellner, B. (1998). "Objective Evaluation of grapheme to phoneme conversion for text-to-speech synthesis in French", in *Computer Speech and Language, Special Issue on Evaluation*, Volume 12, Issue 4., October 1998, 393-410.
32. Zhang, J., Dong, S., and Yu, G. (1998). "Total Quality Evaluation of Speech Synthesis Systems", in *Proceedings of the 5th International Conference on Spoken Language Processing*, Sydney, 30 Nov – 4 Dec 1998. Volume 5, 1711-1714.