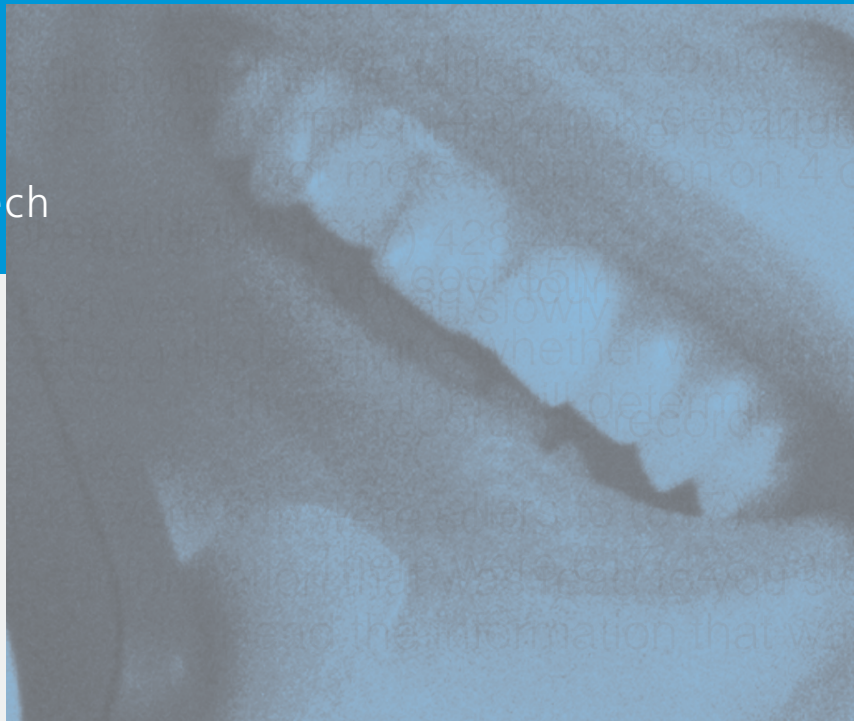


■ *White Paper*

## Assessing Text-to-Speech System Quality



## Introduction

When you create a text-to-speech (TTS) solution, your choice of a TTS system is naturally critical. The aim of this white paper is to assist the reader in understanding how to carry out an effective evaluation and how to interpret evaluation results. The tests are applicable to both concatenative systems (i.e., systems that select and join together segments of real speech from a database to read out a text, like RealSpeak™ Telecom and RealSpeak Solo) and formant systems (i.e., systems that synthesize the speech signal from scratch, like SpeechWorks Solutions' ETI-Eloquence™).

Below are the main contributing factors that SpeechWorks Solutions believes affect the overall quality of a TTS system.

**Intelligibility** – how much of the spoken output can you understand? How quickly do you become fatigued as a listener?

**Naturalness** – how close to human speech does the output of the TTS system sound?

**Front-end processing** – the ability of the system to deal intelligently with commonly used challenges in text such as abbreviations, numerical sequences, homographs and so on.

*That cost \$5M.*

*I'll record the record.*

*There were 617,428 callers to (617) 428-4444*

Having defined these three categories, it's clear that they are not independent of one another; serious front-end errors will lead to less intelligible speech, which will be perceived as less natural. Although no single evaluation provides a foolproof answer, a thorough assessment that focuses on these three issues should provide you with useful and reliable data to help you select the right TTS system for your needs.

***Throughout the paper, entered in bold italics, we also suggest some golden rules to follow when conducting your evaluations.***

It is worth highlighting the fact that evaluation of TTS systems is a notoriously difficult problem. After years of discussion and research in the academic community, there is still no

convincing solution. As a result, assessment remains frustratingly subjective—how a business decides to buy one system over another often hinges on just a few people's opinions based on very limited usage. This is somewhat risky.

It is clear that different listeners respond to different elements of the speech; some are most influenced by intelligibility; some by how human the signal sounds; some by how well the text analysis performs; others by the price. Based on SpeechWorks Solutions' work with dozens of deployed TTS applications, and an analysis of the best-known evaluation methods today, this paper presents our guidelines to effectively evaluate TTS systems.

There are, of course, additional system-related factors that influence business buyers' choices of TTS system for any given application. These include sizing, density and performance issues, ease of use of the software interfaces, platform support, tools support, and so on. This document is focused on evaluating the quality of the TTS output only, and does not cover system-related issues in any detail.

## Before the Evaluation – Practical Considerations

In a perfect world, there would be a simple metric that we could apply to any TTS system to reveal some sort of magic number that describes its overall quality. That metric does not exist. One reason for this is that experts don't really want to assess TTS systems in isolation; it is more useful to assess the performance within a specific application or application type. This abstraction from the raw capabilities of the system makes the notion of a uniformly-applicable magic number less tractable. A second reason the metric does not exist is that factors independent of TTS quality come into play when deciding which system to purchase. Decisions are influenced by many pragmatic factors that are not necessarily related to the technology. These factors include:

**Application type** – Different applications have differing needs for a TTS system. A voice that may be appropriate for a system intended to serve entertaining commercial applications (selling electronic games to teenagers, for example) probably won't be the best for a mainstream banking application.

**Language/Accent** – If you want to deploy a TTS application that speaks Indian English, should you evaluate systems that speak with different accents (e.g., US English, UK English, and Australian English) or should you only consider TTS systems that support Indian English? The answer depends on details of the market segment for which you are deploying the system, but the answer is usually that you should localize to your target market if possible.

**Mixing TTS with Recordings** – For many applications, it may be appropriate to mix TTS with recorded human voices. If so, how well this can be done with a given TTS system should be an important part of the evaluation. For example, the SSML audio tag provides support to seamlessly blend TTS with recorded prompts, which of course should use the same voice talent as the TTS system (cf. <http://www.w3.org/TR/speech-synthesis> )

If we follow this path—how to take into account the various factors that may influence how we would evaluate TTS systems under different circumstances—it may seem that there are too many variables to consider.

It's a difficult problem, but with many TTS systems available, at many different prices, the first thing you need to do is remove as many of those pragmatic variables from your options before you carry out the evaluation. In other words, you should put yourself in a position where you're comparing apples-to-apples.

***Level the playing field. Consider only the systems that meet all your pragmatic requirements; for example, price, disk usage, memory, language, accent, and gender.***

The next question to address is this: how will your application be deployed? If it is going to read email over the telephone, then your evaluation should take place with the subjects listening to samples over the phone. Some systems sound great over speakers, but degrade significantly over the telephone network. When preparing the TTS samples saved as .wav files for the evaluation, you should also ensure that the texts are presented to the candidate TTS systems just like they would be in the deployed application. That doesn't mean you have to have the application fully written, but the texts should look similar to how they'll look in the final version.

The content of the text should also match as closely as possible the application that you intend to deploy. Different TTS systems have differing abilities to read different types of text (e.g., number sequences, dates and times, names, news-style text, email content, etc.). As an example, SpeechWorks Solutions encountered a developer testing systems on a passage from a favorite children's story when the application was developed to read street addresses. The test domain was such a long way from the application domain that the results were inherently unreliable.

Likewise, your test subjects should represent your expected user population. If your customers are aged 18-25, it is best to test your application on such a demographic.

***Match your evaluation environment to your application environment. Different environments yield different results.***

### **Types of Tests to Consider**

In the introduction, we proposed three main criteria for assessing overall system quality: intelligibility, naturalness, and front-end processing. In this section, we discuss tests that help you address these three issues. Note that these factors do affect each other—it's very difficult to completely isolate intelligibility from front-end processing, for example.

#### **Intelligibility**

Intelligibility tests are concerned with the ability to identify what was spoken or synthesized. They are less concerned with the naturalness of the signal, although naturalness inevitably is related to, and influences, intelligibility. Many of these tests are described below. For evaluating and comparing TTS systems for use in commercial applications, SpeechWorks Solutions recommends using a comprehension task because it represents the broadest picture of TTS success.

#### **Comprehension Task**

Of key importance in a commercial environment is whether callers understood what they heard. A common accusation leveled at lower quality TTS systems is that even though you can follow what is being said, it is extremely difficult to access the meaning of what is being spoken. Good TTS must present text in a manner such that the listener easily and correctly comprehends what is

being said. We have experienced cases where people are listening so intently to the signal to understand what the literal sequence of words is, that they aren't paying any attention at all to what the words mean. When asked what the text was about, they had no recollection of any detail.

Our recommendation is to synthesize passages of application text (e.g., emails) and play them to subjects. Then ask them questions about the passage—who was the message from? When did it arrive? What was the main point of the message? What did the message author want you to do?

### Phonetic Tasks

Phonetic tasks deal with identifying the precise sounds within a specific word that are synthesized. They tend not to deal with intelligibility of whole words or phrases or sentences.

**Diagnostic Rhyme Test (DRT)** – DRT is a test of the intelligibility of word-initial consonants. Subjects are played pairs of words with a different first consonant, and asked to identify which word they heard (e.g., dense vs. tense). The system that performs best is the one that elicits the lowest error rate.

**Modified Rhyme Test (MRT)** – MRT is like DRT, but includes tests for word-final intelligibility (e.g., bath vs. bass).

Other phonetic tasks you may find that tackle similar issues to DRT and MRT are: Standard Segmental Test, Cluster Identification Test, and Diagnostic Medial Consonant Test. All these tests require that subjects identify specific sounds within a signal. We suggest that these are useful tools in the development of some synthesizers, but are less useful in practical evaluations of released TTS systems. This is because, in commercial speech services, callers rarely need to identify a single letter in a specific word. Most likely the context and flow of a dialog or an utterance will be a huge aid to intelligibility in any case.

### Transcription Task

**Semantically Unpredictable Sentence Test (SUS)** – Subjects are asked to transcribe sentences that have no inherent

meaning or context, and therefore minimize the possibility of deriving phonetic information from any source but the speech signal itself, e.g.,

*Green ideas sleep furiously*

This classic example of a SUS clearly means nothing, so if you played this to someone and asked him to write down what they'd heard, you would be able to achieve some level of understanding about how intelligible the signal was.

### Naturalness

**Mean Opinion Score (MOS)** – MOS is a well-known subjective scoring method. Listeners are asked to rate the speech quality of different systems, usually synthesizing the same set of sentences.

Typically, subjects are asked to rate the naturalness of the synthesis on a scale of 1-5, where 1 is very poor and 5 is completely natural. All results are summed, and a mean score between 1 and 5 is derived, which is meant to represent the overall naturalness rating for the system.

If conducted carefully, these experiments can yield reliable data. Relative MOS scores from within the same test certainly indicate preferences between systems. However, MOS scores from different systems using different subject pools or synthesized texts cannot be compared since subjects tend to adjust their scores to the range of quality that they are hearing.

In other words, the experimental conditions themselves affect the MOS scores (e.g., whether the subjects listen over headphones or speakers or telephones, the quality of the headphones or speakers, the acoustics in the room, the demographics of the subject population...). In fact, unless you are presented a full set of MOS scores for different engines from the same test, MOS scores will not be able to give you a useful indication of relative performance of the engines.

MOS scores are also highly susceptible to the size of the subject pool, so it's important to use a large enough pool (at least 10 subjects covering 800-1000 sentences for each TTS system) so that outlying results do not unduly affect the overall picture.

**MOS test results are very sensitive to test conditions, the pool of subjects, and the test materials. Be sure to structure your test appropriately.**

It is also important that the tests are done *blind* meaning that the subjects don't know the source of the speech their hearing, so as to avoid any bias towards or against any of the systems. SpeechWorks Solutions uses automated test tools, so the tests are inherently *double-blind* (i.e., neither the subject nor the test invigilator can influence the output).

All the examples from each TTS systems should be played to the subject set without revealing the identity of the systems. Sentence and system order should be randomized so that the subject is less likely to develop a bias during the test (always preferring the second system for example).

For example, if we want to test system A, B and C, we might have them synthesize a set of 3 sentences (note that in real-life you would want to use more sentences—around 50-100 per subject). We would not present the speech files in this order:

- Sentence 1, version A*
- Sentence 1, version B*
- Sentence 1, version C*
- Sentence 2, version A*
- Sentence 2, version B*
- Sentence 2, version C*
- Sentence 3, version A*
- Sentence 3, version B*
- Sentence 3, version C*

This illustrates a case where we would expect order effects: by the time the subject reaches version C of each of these sentences, there is no new information to hear, which makes comprehension significantly easier. This is likely to lead to unduly positive ratings of system C. If we randomize the system order for each sentence, we reduce order effects.

- Sentence 1, version C*
- Sentence 1, version B*
- Sentence 1, version A*
- Sentence 2, version A*
- Sentence 2, version C*
- Sentence 2, version B*

- Sentence 3, version B*
- Sentence 3, version A*
- Sentence 3, version C*

Alternatively, we can completely randomize the order of every item, as below. This is beneficial because many subjects tend to exaggerate score differences when different versions of the same sentence are played side by side. This randomization is not as important as the randomizing system order. SpeechWorks Solutions has carried out all types of test, and find that as long as system order is randomized, the trends are repeatable and results are reliable.

- Sentence 3, version C*
- Sentence 1, version B*
- Sentence 1, version A*
- Sentence 2, version A*
- Sentence 1, version C*
- Sentence 3, version B*
- Sentence 2, version B*
- Sentence 3, version A*
- Sentence 2, version C*

Our recommendation is to carry out a carefully designed MOS test, using no fewer than 10 subjects, each listening to no fewer than 50 items. Randomize the order in which different versions of the sentences are presented, and ensure that overall, each subject hears the same set of sentences from each system.

**Forced-Choice Ranking** – It is rare that we would want to assess a TTS system without comparing it to another system. It is easier—and more practical—to say, “A is better/worse than B” than it is to independently say, “A is good” or “A is bad”. While forced-choice ranking is not an especially useful diagnostic tool, it can be useful as a means to show us the relative overall perceptions of a set of systems.

In this type of test, subjects are played different TTS systems' renditions of the same set of sentences, and asked to rank the renditions of each sentence. Given a large enough sentence set, and a large enough set of subjects, useful information can be derived about how frequently each system is ranked top, first or second, in the top half, and so on. However, we do not recommend this type of test above a well-designed MOS test.

## Front-end processing

**Functionality Test** – This is the closest to a truly objective test. For this test, you should gather sentences that contain multiple examples—in multiple contexts—of the types of text anomaly that you expect your application to encounter. Synthesize the sentences on each TTS system you are considering, and simply mark as correct or incorrect the wording of the output. You may also want to weight certain problems according to how common they are, or how important they are to your application.

One noteworthy point: capturing multiple contexts for text analysis evaluation is extremely important. To illustrate, consider some of the ways to expand the symbol ‘-’ below;

**My social security number is 345-23-9803.**

*My social security number is three four five two three nine eight zero three.*

**From 1975-1979, I was studying at the university.**

*From nineteen seventy five to nineteen seventy nine, I was studying at the university.*

**Our telephone number is (607) 266-7025.**

*Our telephone number is six owe seven, two six six, seven owe two five.*

If your evaluation doesn’t contain multiple cases, a system that always expands the hyphen symbol to the word “to” might come out ranking artificially well if that is the only context you test. The goal of text normalization is to strike the right balance between getting things right and not over-generalizing (e.g., incorrectly expanding the hyphen to the word “to” all the time).

Our recommendation is to use a Functionality Test and include the broadest set of samples that matter to your evaluation.

## Other TTS Tests

**Perceptual Evaluation of Speech Quality (PESQ)** – This is a tool that predicts subjective reactions to distortions of a signal, based on a priori knowledge of the results of large subjective tests. It is often used to measure degradations of a speech signal over a telephone network.

## Perceptual Speech Quality Measurement (PSQM) –

PSQM also measures the degradation of a signal over a telephone network. PSQM is based on a psycho-acoustic model of human perception. Advocates of PSQM claim a correlation between objective PSQM scores and subjective MOS scores. We find that MOS tests are more applicable to TTS evaluations than these tests because we can more directly see the effects of different types of text input on listeners’ subjective evaluations. In addition, we can tailor MOS tests towards different demographics depending on application types, and this yields different and more precise results.

## Good Practice – How to Carry out the Tests

Whatever evaluation approach you select, here are suggestions on how to administer the evaluations in order to obtain the most accurate and useful test results:

- Use a large enough pool of subjects to account for outliers. Different individuals often disagree enormously in TTS assessment, so a large subject set is needed to expose genuine trends. As in many such studies, more is better, so large enough depends on your budget for the assessment, but we find that 10 subjects listening to 50-100 sentences each from each of the systems provides a good balance between cost and reliability of the results.
- Select subjects that emulate the target market’s population distribution as much as you can.
- Make the test blind (where the subjects are unaware of which systems are being evaluated and in what order) or double blind (where both the subjects and the test administrator are unaware of which systems are being tested). Subjects are frequently sensitive to priming. Previous opinions about a known system may affect current feedback. Do not, of course, use subjects who are already familiar with one or more of the TTS systems since they are likely to be able to recognize the system identity from the voice.
- Provide a simple and specific feedback form. It is much easier to derive statistics from answers to direct questions and 1-to-5 rankings than from freeform feedback.

However, because freeform feedback can be useful, it's worth leaving time and room for additional comments at some points in the test.

- Don't get too technical with your questions. Naïve listeners are generally preferable for evaluations as they represent the target audience. Asking expert-level questions about phonetic accuracy, durations, intonation, etc. usually leads to unclear and contradictory answers.
- In a large-scale evaluation, it is not desirable to ask TTS experts for feedback. They tend to respond very differently than naïve listeners (focusing on known problems with TTS systems rather than overall impressions, for example), and don't usually give representative results. For the front-end processing evaluation, it makes sense to have a vendor-independent expert (someone with a linguistics background) devise the test materials, and to decide what outputs are acceptable. This will allow you to score the outputs more objectively.

### How does SpeechWorks Solutions evaluate TTS systems?

SpeechWorks Solutions has experience with most of the tests described here. To help you devise your TTS evaluation plan, let us briefly discuss why we select not to use certain methods for overall assessment of TTS systems.

DRT (Diagnostic Rhyme Tests), MRT (Modified Rhyme tests), and related detailed intelligibility tests are useful in development, but very limited for assessing the quality of a TTS system within an application. Experience has proven that it is more effective to evaluate TTS systems under similar conditions to real-use—reading text from likely TTS applications.

In addition, we do not use PSQM or PESQ, because their primary function is to assess signal degradation over telephone networks.

SpeechWorks Solutions has used forced-choice rankings in the past, and they give useful, but crude results. Today we recommend using carefully designed and administered relative MOS tests.

We have carried out front-end functionality tests of the type described above also, and found them to be an effective means of assessing some aspects of the front-end quality.

### Summary

SpeechWorks Solutions considers three areas to be critical to a thorough evaluation of TTS systems:

- Intelligibility
- Naturalness
- Quality of front-end processing

We've discussed many evaluation methods in this white paper. Some of these tests were designed as diagnostic aids to help TTS developers improve their systems. We recommend the tests that are more suitable for making decisions about which TTS system to use for a given application. In summary, these are:

- A relative MOS test to measure naturalness and overall perception of quality
- A comprehension test to measure the intelligibility of the systems
- An objective measure of the correctness of the text processing in the front-end of the TTS system.

If your speech service requires TTS in multiple languages, we recommend performing these tests across the range of languages and voices needed for a given application. It is also very importance to conduct all of these tests within the context of the desired application, so that the text and subjects match as closely as possible the intended use and audience of the system.



SpeechWorks Solutions  
Division  
ScanSoft, Inc.  
9 Centennial Drive  
Peabody, MA 01960  
www.ScanSoft.com

© 2004 ScanSoft, Inc. All rights reserved. ScanSoft and the ScanSoft logo, SpeechWorks, OpenSpeech, DialogModules, RealSpeak, SpeechPAK, and SpeechSecure are registered trademarks or trademarks of ScanSoft, Inc. in the United States and other countries. All other company or product names may be the trademarks of their respective owners.