Text-to-Speech Accuracy Testing - 2003

The information provided in this report presents the accuracy performance of the commercially available Text-to-Speech (TTS) products. Using a test corpus of over one thousand phrases each unit was tested, scored and the results tabulated. Over 20 TTS products were tested. Included were products from: Aculab; AT&T, Babel Technologies; Cepstral; Elan Speech; Fonix; IBM; Loquendo; Nuance; Microsoft; Rhetorical Systems; ScanSoft; SpeechWorks; Voiceware (NeoSpeech); and Winbond. The test corpus was segmented into categories which included: Number Handling; Homograph Handling; Handling of Words of Foreign Origin; Acronym Handling; Abbreviation Handling; Name Handling; and Address Handling and the test results were presented in these categories. The average score for all units tested was 66.2% correct. Overall scores ranged from 51.2% to 86.7% correct. Within a category, scores ranged from a low of 26.8% to a high of 98.3%. The products that were tested represented the latest available release of their product as of July, 2003.

Table 1.1
TTS Accuracy Testing Summary Results - Server-based Products

		AT&T	*Fonix*	*IBM*	*Loquendo*	*Nuance*	*Rhetorical*	*ScanSoft*	*SpeechWorks*	*Voiceware*
		NV	*FAAST*	*IBM TTS*	*Loquendo TTS*	*Vocalizer 3*	*rVoice*	*RealSpeak*	*Speechify*	*Voicetext*
Total Correct	N	688	670	747	776	804	705	700	890	816
	%	67.0	65.2	72.7	75.6	78.3	68.6	68.2	86.7	79.5
Number =	1,027	Total Number =		9,243	Total Correct =		6,796	Avg Correct =		73.5

Table 1.2
TTS Accuracy Testing Summary Results - Embedded/Desktop Products

		Aculab	*Babel*	*Cepstral*	*Elan*	*Fonix*	*Microsoft*	*ScanSoft*	*SpeechWorks*	*Winbond*
		TTS	*Babel*	*TTS*	*Tempo*	*DECtalk*	*TTS*	*TTS3000*	*ETI-Eloquence/ Speechify Solo*	*TTS*
Total Correct	N	573	586	526	555	640	647	700	890	519
	%	55.8	57.0	51.2	54.0	62.3	63.0	68.2	86.7	50.6
Number =	1,027	Total Number =		9,243	Total Correct =		5,635	Avg Correct =		61.0

The testing (as well as the entire development of the test program) was funded entirely by Voice Information Associates. This makes the testing and the results that were obtained unique in that they are completely unbiased and independent of the influence of any of the vendors. They represent the most objective assessment of the relative accuracy of commercially available TTS products that existed as of July, 2003.

A paper that discusses TTS accuracy and the VIA TTS accuracy testing is Text-to-Speech - Naturalness and Accuracy

Table of Contents:

Text-to-Speech Accuracy

VOICE INFORMATION ASSOCIATES

Text-to-Speech - Naturalness and Accuracy

Introduction

Text-to-speech (TTS) technology offers the capability for the conversion of textual information directly into speech. TTS has a number of solid advantages when compared to the alternate which is recording of speech. These advantages include:

1. Significant reduction in speech storage requirements. 500:1 is typical.

2. Significant reduction in transmission bandwidth requirement. 500:1 is typical.

3. Significantly lower production costs.

4. Significantly lower maintenance/support costs.

Despite these advantages, TTS has been used primarily in applications in which either the vocabulary is extremely large and/or it is dynamic. For these applications, the use of recorded speech is not possible/practical. This represents a tiny % of the total telephone-based applications. To date, the potential for TTS has been under-optimized.

A fundamental limiting factor re the broad-based utilization of TTS in speech-enabled applications appears to be the relatively low accuracy of most TTS products. This goes well beyond basic intelligibility. A service that does not provide accurate information is unlikely to be utilized extensively in either a personal or business environment.

The approach of modifying the source text to accommodate the limitations of the TTS product is a reasonable one, given the accuracy limitations existing in most TTS products. Unfortunately, this severely limits the potential TTS market. It also virtually eliminates items 3. and 4. as advantages that TTS technology has.

History

Naturalness limited in Formant Products

Naturalness is the factor that has been the primary evaluation element for TTS products over the last three decades. From the early 1970s until the late 1990s, commercially available TTS products were invariably based on formant synthesis technology where the speech was generated based on a model of the human voice. Naturalness varied considerably from one product to the next. Despite some progress in making them more natural sounding, the basic technology approach limited the ability to achieve speech that could approach something that sounded like speech from a human. The basic sound of the voice varied from one formant synthesis product to the next. Evaluating the sound of these products tended to be highly subjective. Many of them had a number of voices to chose from. In most instances, one of the voices (usually the male) was distinctly superior although they often all had the same basic sound.

Concatenative TTS improved naturalness dramatically

In the late 1990s, TTS products based on concatenative speech synthesis technology appeared on the market. The first was the RealSpeak product introduced by L&H (now ScanSoft). This was a distinct break-through in terms of naturalness. It approached human speech in terms of its naturalness. Admittedly, for certain phrases, the speech would not sound perfectly natural, but still quite a close approximation to what a human would sound like. The broader selection of different voices with truly different sounds was also available. Since then, a number of TTS products based on concatenative technology have been introduced. Aculab, AT&T, Babel Technologies, Cepstral, Elan Speech, Fonix, IBM, Loquendo, Microsoft, Nuance, Loquendo, Rhetorical Systems, SpeechWorks, SVOX, and Voiceware have all introduced TTS products based on concatenative technology.

Naturalness Difficult to distinguish in Concatenative Products

Concatenative TTS products have progressed to the point where the naturalness differences are really not that variable. For units with like configurations (memory size and channels supported) the naturalness test results suggest that little real difference exists. It has been generally found that the relative Mean Opinion Score (MOS) of individual TTS units correlate quite well with the number of channels that are supported, which in turn is proportional to the memory that is required for a particular TTS product. This suggests that the reliance of subjective tests such MOS as a primary method of assessing the relative naturalness of TTS units is not as appropriate today as it was just a couple of years ago. The utilization of concatenative technology has advanced to a point where little real difference exists in the basic naturalness of the units. Any degradation of naturalness

appears to be readily determined by examining the channel capacity specification of the unit, which is considerably more efficient than running extensive MOS tests..

Just how important is naturalness?

The naturalness of TTS is undoubtedly important. Everything else being equal, the TTS product that is the most natural will win. Everything else, however, is rarely the same. Channel capacity is generally inversely proportional to the naturalness. This means that a price that is paid for naturalness is that the price per port is significantly higher. More importantly, though, the strong focus on naturalness has tended to obscure the fact that severe accuracy deficiencies do exist in many TTS products. How important is naturalness to the user? The little actual data that has ever been gathered and analyzed suggests that it is not as important as is generally believed. If the TTS delivery provides the information/service in an intelligible and useful fashion, the relative naturalness of the TTS is not important. The naturalness is only important to the extent that it compromises the information delivery. For most speech-enabled applications, best results are achieved when the TTS voice is transparent to the user.

Despite the effort by many of the leading industry vendors to create user interfaces that have persona and achieve an anthropomorphic model, data suggests that this is inappropriate and will generally not be acceptable to the user. Modeling the voice to behave similar to a live agent is acceptable. Going beyond this by creating a persona that is "cutesy" or overly friendly has been found to run the risk of yielding a severe user backlash. This further reinforces the view that the existing naturalness level of TTS products is quite adequate and that further improvement has little real utility.

Accuracy has not improved very much

Improvement in the accuracy of TTS products has been relatively static during the same period in which the naturalness was improving so dramatically. The basic letter-to-sound rules and techniques for the handling of numbers, acronyms, abbreviations, words of foreign origin, names, etc. are virtually identical to the techniques that were employed two decades ago.

Accuracy is what Counts

Although it is quite important in the TTS selling process, it turns out that naturalness is not as important to the person that ultimately counts the most - the user. The primary intent of TTS is to communicate information accurately. The definition of what is a natural voice tends to be a oriented towards the way that humans speak in conversation. This is how we evaluate them and what the designers of TTS products strive to achieve. In call automation applications, this doesn’t make an awful lot of sense, since the vast majority of the TTS applications are ones in which a conversational speaking style is inappropriate. A person reading text to themselves does not utilize the lively style that is common in conversational speech. E-mail reading is not typically like a poetry reading!

One could argue that reading e-mail in a conversational fashion doesn’t do any harm. The reality is that it does. The liveliness requires considerably more effort on the part of the listener. Fatigue, headaches, as well as driving accidents are a result.

The disproportionate effort on "naturalness" also means that the amount of effort that is spent on addressing accuracy is reduced. A natural-sounding TTS system is not necessarily accurate.

Evaluation of TTS Accuracy[1]

The process of evaluating the accuracy of a TTS product is relatively straightforward. You send test to it and you listen to what is spoken. The criteria for correctness would be what a typical person would say. An assumption is that a person is familiar with the word, and although they may not know the meaning, they have heard it said. This is a generally objective process. If the TTS product says it correctly, they obtain one score. If it does not say it correctly, it is given a lower score. Adding the score on individual words or phrases for each TTS product provides a mechanism for comparing the relative accuracy of different TTS products. The higher scoring units are more accurate. The assessment is quite objective. Any subjectiveness due to the biases/ignorance of the tester that is listening is applied fairly to all of the units being tested. Admittedly the results are based on a sample. If, however, the sample is made large enough, the results provide reasonable insight into the relative accuracy of the TTS products being tested.

TTS Context Analysis

Context analysis in virtually all of the TTS products is quite limited. It is typically confined to utilizing basic rules that are directed at well defined formats/conventions. For example telephone number processing, ZIP code processing, and the handling of many abbreviations (does Dr. mean Drive or Doctor?). Homographs (a frequent occurrence in text) are often not analyzed properly in a contextual fashion. In the VIA testing of 100 of the more common homographs, the average correct score was only 56.8%. Common ambiguities are often not handled properly. Worst of all, they are handled differently in each TTS product.

Given that meeting the expectations of the user is the most important goal of a speech-enabled system and the user is likely to encounter user interfaces implemented with different TTS products, the industry is presently falling somewhat short at meeting this most fundamental user interface requirement.

Consistency of TTS Products

From the perspective of the user, consistent handling from one TTS product to the next is important. Very little consistency existed in the products. Acceptance of TTS (as well as any product/service) is strongly tied to meeting the expectations of the users. This is obviously not being achieved since the user experiences something different with each TTS product. A standardized approach to TTS accuracy would be of benefit to the user and improve the acceptance of TTS.

What needs to be tested?

Providing the "correct" pronunciation of a particular word is difficult in the English language. Some languages (Spanish is an example) have a close "fit" between the orthographic (spelling) system and the phonetic system. The English language does not have this close fit. In fact, learning to read in English (pronounce a word from its written

form) is extremely difficult. Numerous exceptions to the pronunciation rules exist. These exceptions include:

Number handling

Foreign words

Acronyms

Abbreviations

NamesAddresses

Homographs

Punctuation

Incorrect writing style

TTS Products that were Tested

The TTS products that were tested for accuracy were:

Aculab TTS

AT&T Natural Voices

Babel Babel TTS

Cepstral TTS

Elan Tempo

Fonix FAAST

Fonix DECtalk

IBM TTS

Loquendo TTS

Microsoft TTS

Nuance Vocalizer 3

Rhetorical rVoice

ScanSoft RealSpeak

ScanSoft TTS3000SpeechWorks Speechify, SpeechWorks ETI Eloquence

SpeechWorks Solo

SVOX TTS

Voiceware Voicetext

Winbond TTS

Overall TTS Accuracy Summary

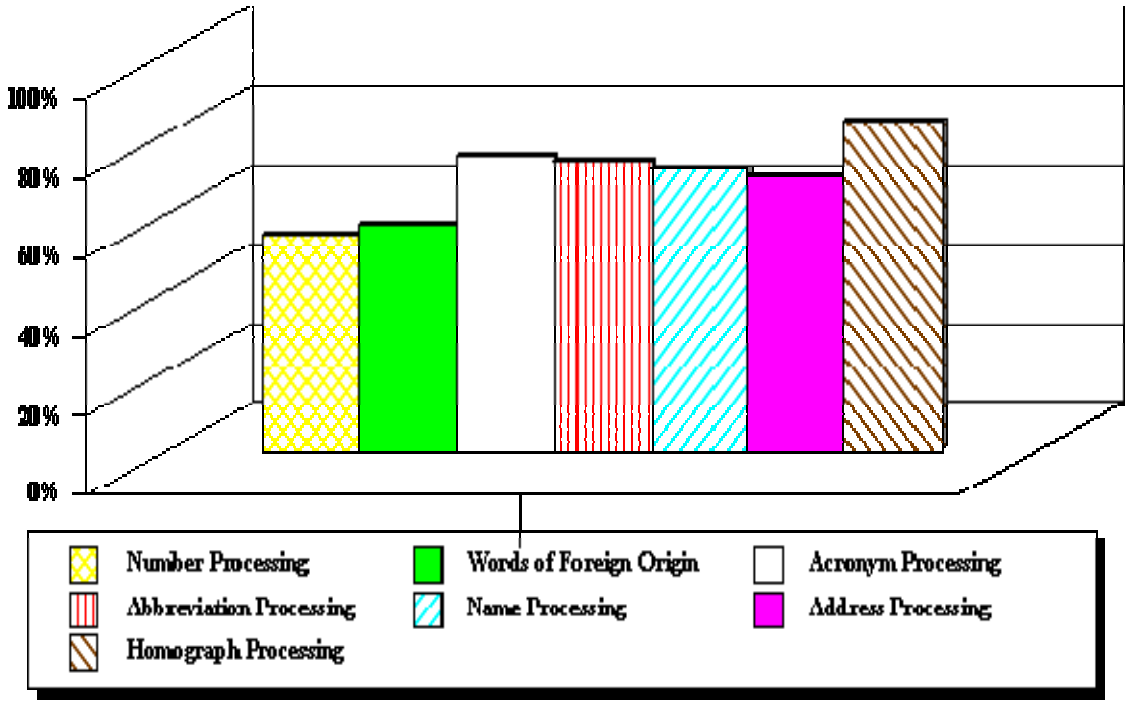
Table 1

TTS Accuracy Summary

TTS Accuracy Test	***Average % Correct***
Number Processing	55.6%
Words of Foreign Origin Processing	58.8%
Acronym Processing	74.1%
Abbreviation Processing	72.9%
Name Processing	70.7%
Address Processing	69.0%
Homograph Processing	83.4%

The areas in which the TTS products were weakest were in Number Processing and in handling Words of Foreign Origin. The other areas were relatively strong, with an error rate of less than 1/3. The TTS products did best with Homograph Processing with an accuracy of 83.4% for the test phrases.

Figure 1 TTS Accuracy Test Results Summary



TTS Number ProcessingHandling numbers incorrectly is one of the most common problems with a TTS product. The most common errors are:A minus (-) is either ignored or it is spoken as "dash".

A > or < is frequently ignored.

A dash separating numbers is either ignored or spoken as "dash"

The slash is handled improperly (1/3 is spoken as one-slash-three or one over three)

Months are mixed up with names (Jan. is either January or Jan.).

Date variants are treated as number strings or arithmetic expressions..

These errors can be serious. In some instances, the information is conveyed incorrectly.

Numbers are commonly encountered. Dates, telephone numbers, addresses, account codes, times, dimensions, weights, volumes, speed, and ranges are among the more common uses of numbers in text.

The TTS Number Processing testing that was done is shown in Table 2. The average % correct was 54.8%. The range was as low as 33.3% correct to as high as 72.7% correct.

Table 2

TTS Number Processing

TTS Unit	***#1***	***#2***	***#3***	***#4***	***#5***	***#6***	***#7***	***#8***	***#9***	***#10***	***#11***	***#12***	***#13***	***#14***	***#15***	***#16***	***#17***	***#18***	***#19***
% correct	48.5	56.9	60.6	54.5	51.5	69.7	60.6	66.7	54.5	69.7	60.6	66.7	72.7	72.7	54.5	54.5	39.4	35.4	33.3
Average	54.8%																		

A number of highly ambiquitous cases exist. For example: How should 9/11 be spoken?**Table 2a**

TTS Number Processing Ambiquity Example

Input Text	***Contextual Phrase***	***Phrase #***	***Output Phrase***
9/11	?	1	nine slash eleven
	?	2	0.81818182
	Nine eleven was a dark day for America	3	**nine eleven**
	The date was 9/11.	4	**September eleventh**
		5	September eleven
	The distance was 9/11 of a mile.	6	nine over eleven
		7	nine divided by eleven
		8	**nine elevenths**

We have eight (8) ways in which 9/11 could be spoken that could be argued are all technically correct. The issue is not whether it is technically correct. The decision as to what should appropriately be spoken should address firstly the question: What would a person say when they saw 9/11? This would eliminate the first two phrases, since a person is highly unlikely to say this. It would also eliminate the 5th, 6th and 7th phrase responses. At this point, we are left with three choices. Without contextual analysis that would determine that 9/11 was a measurement or a date, it is impossible to select a treatment that would be 100% correct. The best choice is now one of the assessing the relative probability of occurrence. The choice is certainly dependent on the domain. Choosing "nine eleven" would appear to be the safest choice. It would be the most correct when referring to the event that occurred on September 11, 2001. It would be less appropriate for 9/11 meaning the general date, but still quite reasonable. When used as a measurement, it would be incorrect but not totally erroneous. The specific example chosen had an extra dimension of ambiquity that would not occur for ½. On the other hand, for ½, the spoken output possibilities are: "one- half', "one-two", "one divided by two", "one slash two", "one over two", "February second", "February two", and 0.5000, which is also eight possibilities. An additional dimension of ambiquity is introduced by some of the TTS products which will pronounce it differently if a minus sign is before it or a % sign is after it. It could certainly be argued that a safe choice is a better basic design approach than one that is perfectly correct 90% of the time and totally wrong 10% of the time. Although this is certainly a

valid view, the choice made by many of the TTS designers appears to be to be incorrect 99% of time and still be totally incorrect 1% of the time.

Proper handling of number strings is most important, particularly for e-mail communications. Numbers separated by a dash (-) were one of the most problematical. The TTS units that didn't handle this properly either totally dropped the dash (-) or performed arithmetic on the number string. A number preceded by a minus (-) or a less than (<) or greater than (>) sign were another problem area for most TTS units. These sorts of errors are not minor. A business tool that is unable to handle numbers properly is not very useful.

TTS Handling of Words of Foreign Origin

A number of commonly used words are ones that were taken from a non-English language. The letter-to-sound rules of the English language simply do not apply. Included were approximately 85 words that are frequently used in the language. Probably used more commonly in conversation than in text, but never-the-less, occurring in text.

Table 3

Words of Foreign Origin

TTS Unit	***#1***	***#2***	***#3***	***#4***	***#5***	***#6***	***#7***	***#8***	***#9***	***#10***	***#11***	***#12***	***#13***	***#14***	***#15***	***#16***	***#17***	***#18***	***#19***
% correct	54.8	63.1	70.2	67.9	60.7	61.9	60.7	73.8	64.3	67.9	58.3	63.1	58.3	58.3	63.1	63.1	54.5	39.4	38.4
Average	57.1%																		

Some level of legitimate debate exists re the correct pronunciation of words of foreign origin. Many people do not pronounce many of them in the way that the test required. The criteria used for testing was the current dictionary definition.***TTS Acronym Handling***

Acronyms are commonly used in the English language. Acronyms differ from abbreviations in the user not typically speaking the non-abbreviated form. For example, users typically read **NASA** as nah-sah and not N A SA or National Aviation and Space Administration. Ambiguities exist. **ASAP** is as likely to be spoken as A-sap or as A S A P or as soon as possible. **radar** is spoken as ray-dar and not as R A D A R or radio detecting and ranging. NATO should be pronounced nay-toe and not **N A T O**. In addition to it meaning North Atlantic Treaty Organization, it could also mean: National Association of Theater Owners, Inc. National Association of Timeshare Owners; North African Theater of Operations; North American Turbocoupe Organization (Ford Thunderbird Turbocoupes). The acronym DEC has multiple possibilities among them: December, Decimal, Digital Equipment Corporation, Data Entry Clerk, Data Equivalence Class, Decade, Decatur, GA, USA (airport code), Decatur, IL, USA - Decatur Municiple Airport (Airport Code), Deceased, Decision, Declaration, Declarations Page (insurance), Declared Declination, Decorated, Decorative, Decrease, Decrement, Decrescendo, Deductible Employee Contribution (US IRS), Delta Executive Controller, Department of Environmental Conservation, Department of Environmental Conservation (State of New York), Design Error Check, Development Education Centre, Development Experience Clearinghouse (USAID), Dévéloppement Économique Canada pour les Régions du Québec, Device Clear, Diethylcarbamazine, Digital Engine Control, Diplôme d'Études Collégiales, Direct Energy Conversion, Dispersion Equalization Card, District Export Council, Division for Early

Childhood (Council for Exceptional Children), Douglas Electric Cooperative This one presents a dilemma: The abbreviation for December is quite common. Pronouncing the letters would appear to be correct for instances other than this one.

The number of acronyms that exist in the English language numbers in the hundreds of thousands. Virtually every acronym is ambiguous in a theoretical sense.

The acronyms that were used in the TTS testing were generally ones that were well known by most people. Couldn't resist ones such as ICASSP, VoIP, WAP, AVIOS and SAPI, though.

One vendor debated that I C A S S P was an acceptable way of handling ICASSP. This is a reasonable view in a theoretical sense but incorrect when judged against the basic test criteria that was used in this testing. The likihood of an individual familiar with the industry pronouncing individual letters is as likely as them saying "International Conference on Acoustics and Speech Signal Processing."

Acronym processing is an area in which domain-specific processing would be highly appropriate, since most of the ambiquities that exist relate to the domain in which an acronym is being used.

Table 4

TTS Acronym Handling

TTS Unit	***#1***	***#2***	***#3***	***#4***	***#5***	***#6***	***#7***	***#8***	***#9***	***#10***	***#11***	***#12***	***#13***	***#14***	***#15***	***#16***	***#17***	***#18***	***#1***
% correct	58.2	69.1	67.3	80.0	81.8	69.1	65.5	76.4	83.6	74.5	74.5	70.9	94.5	94.5	90.9	90.9	69.1	56.4	67
Average	74.2%																		

Over 100 common acronyms were utilized in the testing. These included ones that are correctly pronounced as well as ones that should be spelled.

TTS Abbreviation Handling

The most common errors are with names (Robt., Wm. A., III, Jr., Sr.), addresses (St,. Rd., Hwy., Ln., OR) and punctuation being missing (US vs U.S.).

Table 5

TTS Abbreviation Handling

TTS Unit	***#1***	***#2***	***#3***	***#4***	***#5***	***#6***	***#7***	***#8***	***#9***	***#10***	***#11***	***#12***	***#13***	***#14***	***#15***	***#16***	***#17***	***#18***	***#19***
% correct	57.1	67.9	66.1	78.6	80.4	67.9	64.3	75.0	82.1	73.2	73.2	69.6	92.9	92.9	89.3	89.3	67.9	55.4	66.
Average	72.9%																		

The abbreviations that were used in the TTS accuracy testing were generally ones that were common and readily recognized by the average person. Over 100 common abbreviations were used in the testing.

TTS Name Handling

Proper name handling has to do with the abbreviations that are commonly used and an understanding of language origin of the name. Some subjectivity certainly exists. Individuals have the right to pronounce their own name in whatever way that they wish. Regional variations exist. The criteria that was used in the TTS Name Handling testing was to attempt to assess how a typical person would pronounce the name.

Table 6

TTS Name Handling

TTS Unit	***#1***	***#2***	***#3***	***#4***	***#5***	***#6***	***#7***	***#8***	***#9***	***#10***	***#11***	***#12***	***#13***	***#14***	***#15***	***#16***	***#17***	***#18***	***#19***
% correct	56.1	66.7	64.9	77.2	78.9	66.7	63.2	73.7	80.7	71.9	71.9	68.4	86.0	86.0	84.2	84.2	66.7	54.4	64.9
Average	70.7%																		

TTS Address Handling

Address handling has to do with proper handling of abbreviations and specific address formats/conventions that are generally well defined and readily recognized by the typical human reader.

Table 7

TTS Address Handling

TTS Unit	***#1***	***#2***	***#3***	***#4***	***#5***	***#6***	***#7***	***#8***	***#9***	***#10***	***#11***	***#12***	***#13***	***#14***	***#15***	***#16***	***#17***	***#18***	***#19***
% correct	53.6	64.3	63.4	75.0	76.8	64.3	60.7	71.4	77.7	69.6	69.6	66.1	87.5	87.5	84.8	84.8	64.3	51.8	62.5
Average	69.0%																		

Addresses that were utilized in the TTS address handling testing were selected in a random fashion to obtain a reasonable regional mix.***Homographs***In addition, many commonly used words are homographs. Homographs are words that are pronounced differently depending on their use. If the word is used as a noun it has one pronunciation, while if used as a verb, the pronunciation is different. The "correct" pronunciation is only obtainable by obtaining syntactic information. Over 300 homographs exist in the English language.

Table 8

TTS Homograph Handling

TTS Unit	***#1***	***#2***	***#3***	***#4***	***#5***	***#6***	***#7***	***#8***	***#9***	***#10***	***#11***	***#12***	***#13***	***#14***	***#15***	***#16***	***#17***	***#18***	***#19***
% correct	80.4	85.9	75.0	74.5	71.9	93.8	92.2	89.1	89.1	80.4	79.7	82.8	90.6	90.6	95.3	95.3	81.3	78.4	78.4
Average	83.4%																		

Many of the units handle homographs quite well. We used a sample of sixty-four of the most common homographs in the testing. Trouble homographs included: lead; invalid; bass; conflict; minute; and close.

Writing is Different from Speaking

When people write, they will frequently violate the rules. Some people will send all e-mails upper-case while others will use all lower-case. Punctuation is frequently used as a mechanism for communication. Writers will use the dash liberally to mean: the number following is negative; the following clause is an explanation; this is a compound word;

In the testing it was noted that spaces are usually applied properly to word boundaries, but are inconsistently applied when punctuation is involved.

How can this be improved?

The most fundamental thing that needs to be done is to recognize that people will continue to write e-mails as they have in the past and not change to accommodate TTS reading of e-mails. The second thing is to recognize that most people that receive e-mails are able to read and interpret them correctly (most of the time). For e-mail, a TTS reader should behave in a manner that is as similar as possible to how the recipient would behave. After accepting this as a basic model, the real issues can be dealt with. Admittedly, in a grand sense, abiquities will exist.

Summary

The first priority of TTS products is to provide accurate information. The results of testing sixteen different TTS products suggests that a good deal of room for significant improvement in this area exists. Significant improvements in accuracy are achievable with a minimum of effort by the vendors simply focusing some solid attention on this area. The ambiquities that exist in the English language are a challenge that will require an increase in the intelligence of the TTS products. Context processing appears to be the solution for bringing the TTS products to a point where they are able to approach the ability of a person in reading text.

Other TTS considerations are certainly important

In evaluating a TTS product a number of factors need to be taken into consideration in addition to accuracy. Naturalness is certainly an important consideration. This is generally fairly easy to address. Simply have a number of people listen to a different products is not an unreasonable method. Tests such as Mean Opinion Score (MOS) and Forced Choice (FC) are essentially structured tests that do just this. A variety of intelligibility tests exist that have been developed by the academic community. A number of tests have been developed that address intelligibility. Diagnostic Rhyme Tests (DRT), Modified Rhyme Tests (MRT), Diagnostic Medial Consonant Test (DMCT), and Diagnostic Alliteration

Tests (DALT) deal with the phonetic sounds within words. These tests are useful tools in the development of TTS products, but are of extremely limited value in practical evaluations of commercially available TTS products. The are also quite time consuming (expensive) to run and require significant skill to administer and to analyze the results properly. The result is that TTS products are typically evaluated in a totally subjective fashion by a few users typing some sample text at a TTS product and then deciding on using a particular TTS product based on this rather limited subjective testing. This is similar to a consumer purchasing an auto based only on "kicking the tires" and the auto industry describing the array of testing processes that they utilize to develop the automobile.

TTS Voice Likability

One area that has received little in the way of formal testing is the "likability" of the TTS voice. Users will definitely find one voice more pleasing than another. An evaluation is readily accomplished by using an MOS or FC test with the criteria being graded being how well the listener likes the voice.

TTS Languages Supported

The number of languages that a TTS product supports is important. In some instances, TTS vendors have achieved multiple language coverage by simply packaging different TTS products together. On the surface, these appear to offer multiple languages. Since they are, however, different binaries, they tend to be considerably less efficient and integration/run-time problems frequently surface when they are deployed.

TTS Utilization Within an Application

How the TTS is used within an application is another important consideration. As an example: a leading telephone company was deploying a reverse directory application. They were receiving a large number of complaints from callers that claimed that they could not understand what the TTS was saying. When delivering the requested address to the caller, the implementation mixed recorded speech and TTS. It said "the address that you requested is" with a recorded voice and then used TTS to deliver the actual address that was requested. Caller complaints were virtually eliminated by simply using TTS for the phrase "the address that you requested is". Nothing else was changed. The address information was identical. This appears to be a demonstration of the ability of a caller to adapt to a voice.

TTS Appropriateness for a Specific Domain

Finally, the application domain in which the TTS product is intended to be deployed is an important consideration. If the application is a medical one the language processing requirements will be quite different than for a financial services or automotive navigation application. For some applications, a specialized language processor will be required. Most of the domain specific TTS language processors have been focused on name/address processing. E-mail processors do exist but these focus mostly on the application and not the textual content.

[1] In early 2003, Voice Information Associates ran accuracy tests on each of the leading text-to-speech products that were commercially available. The detailed results are compiled

and presented in a report that is available from Voice Information Associates titled **Text-to-Speech Accuracy Testing - 2003**.