

Analysis Methods for Assessing TTS Intelligibility

H. Timothy Bunnell and Jason Lilley

Center for Pediatric Auditory and Speech Sciences
Alfred I. duPont Hospital for Children, Wilmington DE, USA
and
Department of Linguistics and Cognitive Science
University of Delaware, Newark DE, USA
{bunnell, lilley}@ase1.udel.edu

Abstract

Semantically unpredictable (SU) sentences are often used to assess intelligibility of TTS systems, but analyses of listener responses to SU sentences can be a labor-intensive process. In this paper we compare several approaches to the analysis of data from an SUS task. Data from a study comparing five TTS systems were analyzed in a variety of ways ranging from string edit measures based on carefully hand-corrected phonetically transcribed responses to largely uncorrected words- or sentences-correct measures. Results suggest that a simple sentences-correct measure is adequate when only rank order information is of interest. However, the sentences-correct measure masks the magnitude of differences between systems and should be avoided when it is important to gauge how large the difference in intelligibility is between systems. In preparing response data for analysis, careful human interpretation of listener response data can lead to higher intelligibility measures overall, but does not interact with TTS system or other factors and consequently does not lead to different conclusions when comparing multiple TTS systems. This suggests that largely automated scoring procedures are feasible.

1. Introduction

The use of syntactically well-formed but semantically anomalous sentences in assessing TTS systems was first described in [1]. More recently, [2] describe procedures for generating semantically unpredictable sentence (SUS) materials for evaluating TTS intelligibility. In [2], based on an earlier study ([3]), the recommended analysis procedure is to score whole sentences as either correct or incorrect, requiring every word of the sentence to be correct and in the correct sequence for a sentence to be scored as correct.

While the scoring procedure recommended in [2] is simple to implement, it is very strict (leading to generally lower measures of intelligibility for any given TTS system), and relatively coarse. Concerns with such a coarse measure include the opposing possibilities that it may either (a) mask relatively serious differences, or (b) amplify relatively subtle differences between two TTS systems, making one system appear to be much more or less intelligible than another. The former could happen, for example, when comparing a TTS system that makes about 2 errors per sentence to a system that makes 10 errors per sentence. The latter could happen when comparing a TTS system that makes phonetically subtle errors with relatively higher frequency to a system that makes gross pronunciation errors with somewhat lower frequency. In such cases, it is possible that scoring responses at a more fine-

grained level would yield different intelligibility rankings of TTS systems, or would provide a more accurate measure of the differences between systems than would responses scored at the sentence level.

A second and more general concern related to analysis of SUS response data is the question of how to interpret ambiguous response data, and whether interpretation of ambiguous data influences conclusions to be drawn from a study using SUS material. This is particularly an issue when, as in the study described here, listeners respond by typing their responses into a computer. Typed responses contain a variety of errors. Some, such as simple typos and spelling errors, are of little interest and are presumed to be randomly distributed with respect to the synthesizers being compared. However, such errors may result in responses that are probably correct to be scored as incorrect. If using a between-subjects design, differences in the spelling or typing ability of subjects across groups could artificially increase or decrease real differences between TTS systems. On the other hand, other errors, such as attempts to “phonetically” gloss tokens perceived as non-words, are indicative of real intelligibility problems and may yield valuable information about the strengths and weaknesses of the system under study.

In the following, we explore the consequences of some of these factors on a set of data collected to compare five synthesis systems.

2. Method

2.1. Dataset

The perception experiment in which the current data set was collected was briefly described in [4]. The study was intended to compare the intelligibility of a new TTS system to four existing commercially available systems. A more complete description of that study is in preparation. Here we outline the overall study design to lay out the structure of the data collected.

2.1.1. Subjects

The subjects were 30 University of Delaware students who received a \$10 gift card to a local bookstore in exchange for participation. All listeners were native speakers of American English and reported having normal hearing.

2.1.2. Stimuli

The stimuli were 100 SU sentences generated by each of the five TTS systems. Since this report is not concerned with the

specifics of the TTS systems, they will be referred to simply as systems A, B, C, D, and E.

Per recommendations in [2] the SU sentences were constructed using words of minimal length (all one-syllable) within five distinct sentence frames. Examples of sentences generated for each of the five syntactic frames are shown in Table 1.

Table 1. Examples of each sentence frame. Words in italics are randomly assigned within the frame represented by words in normal font.

FRAME	EXAMPLE SENTENCE
1	The <i>trip</i> <i>talked</i> in the <i>old</i> <i>stage</i> .
2	The <i>state</i> <i>spared</i> the <i>claim</i> that <i>wept</i> .
3	The <i>thin</i> <i>aid</i> <i>brushed</i> the <i>part</i> .
4	Why does the <i>strength</i> <i>trust</i> the <i>dark</i> <i>sound</i> ?
5	<i>Waste</i> the <i>shape</i> or the <i>hand</i> .

Synthetic renderings of all sentences were generated by each of the five TTS systems. Because all five synthesizers were sufficiently SAPI-compliant to be installed on the same Windows computer, sentences were generated directly to waveform files for storage and later presentation.

To reduce the possible effect of amplitude differences inherent to the five synthesizers, all synthetic speech files were adjusted to 72.0 dB RMS amplitude (calculated over the entire synthetic speech file). For all synthesizers, speaking rate and average F0 were left at default levels (in some cases, these were not adjustable). While consistent differences in speaking rate (as measured by raw waveform duration) existed between synthesizers, the differences were not perceptually prominent. The overall average sentence duration was 2.2 seconds and varied from 1.9 seconds (system B) to 2.4 seconds (system E).

2.1.3. Procedure

The five hundred synthetic sentences (100 sentences by 5 synthesizers) were split into five sets for presentation. Each set contained 20 sentences of each syntactic frame. Of the 20 sentences of each frame, four sentences were produced by each of the five synthesizers. This blocking ensured that each trial set of 100 sentences contained an equal number of sentences of each syntactic frame produced by each synthesizer without any duplication of sentences. Each listener was assigned to one sentence set and, hence, was never presented with the same sentence twice. In all, six listeners were assigned to each of the five sentence sets.

Listeners wore headphones and were seated at a computer in a quiet room. After hearing a sentence one time, listeners were provided as much time as necessary to type the sentence as they understood it. When finished, listeners clicked a *Next* button to initiate the next trial.

2.2. Data Reanalysis

The originally reported analysis of data from this experiment was based on the number of content words correct in each sentence [4]. For the present analysis, we have reanalyzed all data using measures based on the edit distance between the listener responses and the original sentences. The edit distance between two strings is defined as the minimum total "cost" of

transforming one string into the other using insertion, deletion, and substitution operations, each operation being associated with its own cost. For the present analysis, the costs of all operations were set to one, so that the edit distance is simply the total number of insertions, deletions, and substitutions. The same general measure can be used whether the strings being compared are strings of discrete word tokens or discrete phone tokens.

The various edit distance measures reported are based on both word-level and phone-level measures. In both cases, we started by designing a dictionary to map both stimuli and listener input onto response tokens. The raw listener input (typed sentences) and stimulus sentences were first tokenized into a set of input word tokens, where a word token is defined as an uninterrupted sequence of alphabetic characters or apostrophes.

For word-level analyses, the response tokens in the dictionary were usually exactly the same as the input tokens, but in some cases, the dictionary would map multiple possible input tokens onto a single response token. For instance, the input tokens *rows*, *rose*, and *roze* were all mapped to the token *rose* (the word form given to the TTS systems in generating the sentence).

For phone-level analyses, a similar dictionary was used to map input tokens onto strings of phonetic symbols, including a word boundary symbol. For this experiment, two phone-level dictionaries were created, an "uncorrected" one and a "hand-corrected" one. The former was generated by running the tokenized listener input through the letter-to-sound rules of one of the TTS systems with a bare minimum of additional hand editing. For example, the default TTS pronunciation of nonce forms was used unless the system chose to spell out the form. In the latter case, an experimenter-supplied pronunciation was used. The "hand-corrected" dictionary was created by further editing the "uncorrected" dictionary, and attempting to interpret the intention of the respondent. For example, all nonce forms were corrected if the automatic transcription did not agree with the experimenter's interpretation of what the listener intended. When the respondent's intention could not be determined completely, a transcription that would result in the best match between stimulus and response was used.

In addition, strings of phonetically uninterpretable listener input (e.g., a string such as ?????) were mapped onto a word boundary symbol with no other phonetic content. This approach allowed us to retain word boundary location information to the extent that it was recoverable from the response data.

Once all sentences were mapped, the word-level and phone-level edit distances between each pair of stimulus and response sentences were computed. For the phone-level edit distances, we chose to disregard several phonetic differences that were represented in the TTS symbol set we used. Specifically, the difference between front and back schwa was ignored, as was the distinction between a flapped /d/ and either an unaspirated /t/ or a full /d/. We also disregarded the distinction between aspirated and unaspirated voiceless stops. Additionally, in computing edit distances, word boundaries could be inserted or deleted, but they could not be substituted with other segments.

Finally, a sentence-level error score (1 = incorrect; 0 = correct) was also computed for each response sentence. A response sentence was scored correct only if the phonetic edit distance between it and the stimulus sentence was zero.

3. Results

To analyze the data, scores were derived by summing edit distances or sentence errors over the four sentences of each frame type from each synthesizer per listener. This resulted in 25 scores (5 frames X 5 synthesizers) per subject that we treated as a completely within-subjects design. Preliminary analyses revealed that the between-subjects factor SENTENCE SET (as described in 2.1.3) was not significant, and consequently it will not be discussed here.

So that sentence, word, and phone-level data were comparable, raw scores for each level were divided by the number of sentences, words, or phones within the sentences from which the score was derived. The resulting proportion data were highly non-normal in their distribution and were consequently arcsine transformed to improve their suitability for analysis of variance. All analyses of variance described below were conducted using the arcsine transformed data; however, only the original proportions are presented in figures.

3.1. Sentence-level scoring

Sentence-level scoring produced results that closely resembled those originally reported in [4]. Overall, the main effect of

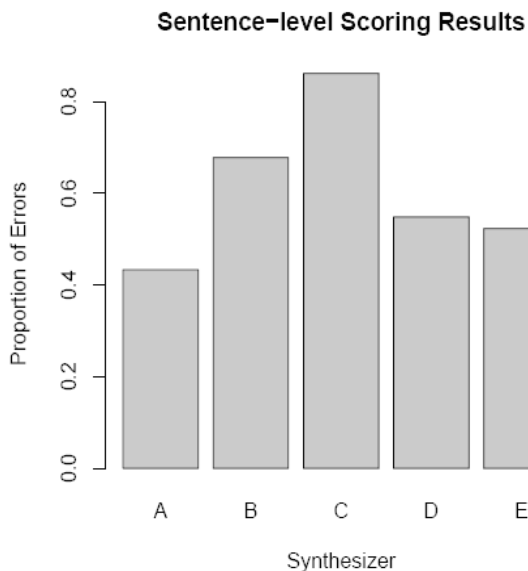


Figure 1: Overall ranking of synthesizer intelligibility when scored at the sentence level.

synthesizer was significant ($F[4,116] = 72.15, p < .001$) as was the effect of sentence frame ($F[4,116]=22.77, p < .001$) and the interaction of synthesizer with sentence frame ($F[16,464]=3.79, p < .001$). Figure 1 displays the means underlying the significant main effect of synthesizer. System A clearly has the lowest error rate and system C the highest. Post hoc tests reveal that all differences among synthesizers except for the difference between systems D & E are significant.

This main effect of synthesizer was conditioned by a significant interaction with sentence frame, indicating that the relative ranking of synthesizers varied significantly over the various syntactic frames used in the study. Figure 2 illustrates this effect by plotting the individual synthesis systems as groups of bars within each sentence frame. As this figure

shows, error rates tended to be lowest for frame 5 (the shortest frame) and are most representative of the overall results. System C has the highest error rate in all frames, and system A

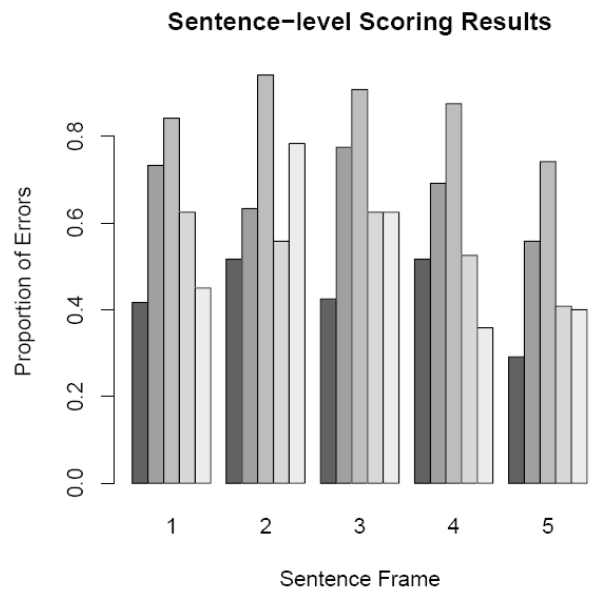


Figure 2: Means for interaction of synthesizer and sentence frame for sentence-level scoring. The shaded bars in each grouping represent the has the lowest error rate in all but one sentence frame. Systems B, D, and E tended to vary more in intelligibility as a function of the sentence frame.

3.2. Word-level scoring

The pattern of significant effects for the word-level analysis of variance mirrored the pattern for the sentence-level analysis with effects for synthesizer, sentence frame, and their interaction all significant ($F[4,116]=161.25; F[4,116]=42.63$; and $F[16,464]=4.05$ respectively, all p 's $< .001$).

The means underlying these significant effects also patterned similarly to those from the sentence-level analysis. As with the sentence-level analysis, system A had the lowest overall error rate and system C the highest (see Figure 5). Systems D and E remained statistically equivalent, although error rates were slightly higher for system E than for system D, a reversal of the order seen in the sentence-level analysis. Two other differences are worth noting. First, as expected, there was an overall lower proportion of errors for all systems. Overall, 60.9% of the sentences in the sentence-level analysis contained errors. However, only 17.9% of the words were incorrectly identified in listener responses. Another noteworthy difference between the sentence-level and word-level analyses is revealed by considering the magnitude of the difference between the synthesizer with the highest error rate and the system with the lowest error rate. For the sentence-level analysis, system C had an error rate (86.2% of the listeners' response sentences had errors) about twice the magnitude of the error rate for system A (43.3% of the listener responses to sentence from system A had errors). By contrast, for the word-level analysis, the error rate for responses to system C (34.6%) was more than three times that of responses to system A (10.1%).

3.3. Phone-level scoring

We turn next to results from analysis of the phonetic edit distance measure. For this analysis we used the edit distances obtained using the largely uncorrected dictionary. Once again, the main effect of synthesizer was significant ($F[4,116]=155.35, p < .001$) as was the effect of sentence frame ($F[4,116]=23.85, p < .001$) and the interaction of sentence frame with synthesizer ($F[16, 464]=3.30, p < .001$).

Figure 3 shows the means underlying the main effect of synthesizer. Comparing this to Figure 1, it is clear that the relative ranking of TTS intelligibility is unchanged. However, the differences between the poorest and best systems are enhanced by using the edit distance measure. Thus, while Figure 1 shows that slightly more than 40% of the sentences for system A contained some error, Figure 3 shows that on average, these were due to errors on only about 4% of the phonetic segments within those sentences. Also of note once again is the relative number of errors on system C versus system A. At the phone level, listeners made more than 4 times as many errors transcribing utterances for system C compared to system A.

Phone-level Edit Distance Results

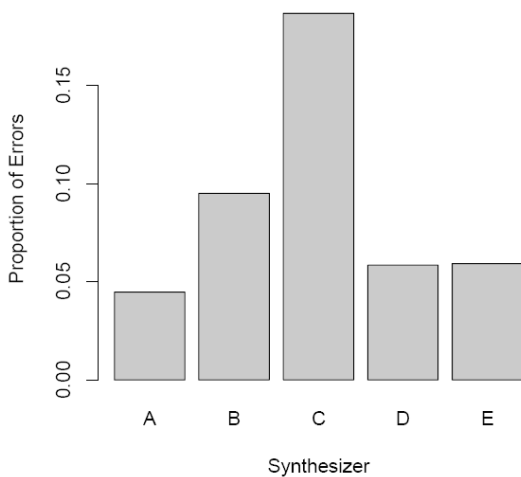


Figure 3: Overall ranking of synthesizer intelligibility when measured as phone-level edit distance.

The significant interaction between synthesizer and sentence frame for phonetic edit distance is illustrated in Figure 4. By comparison to Figure 2, we can see that differences between the systems are generally enhanced. While system C has the highest error rate in all sentence frames, there is greater variability among the other systems in Figure 4. For instance, in simple rank order, system A has the lowest error rate in 3 of the 5 frames which system E has the lowest rate in 2 of the five. As in the sentence-level analysis, however, system A is least variable over the five sentence frames.

3.4. Combined multi-level analysis

To further verify the impression that results from analyses at each level of analysis are qualitatively different, an additional analysis of variance was calculated combining data from all three levels of analysis as an additional within-subjects factor. Results of this analysis are given in Table 2 and Figure 5. As the ANOVA results shown in Table 2 indicate, level of

Phone-level Edit Distance Results

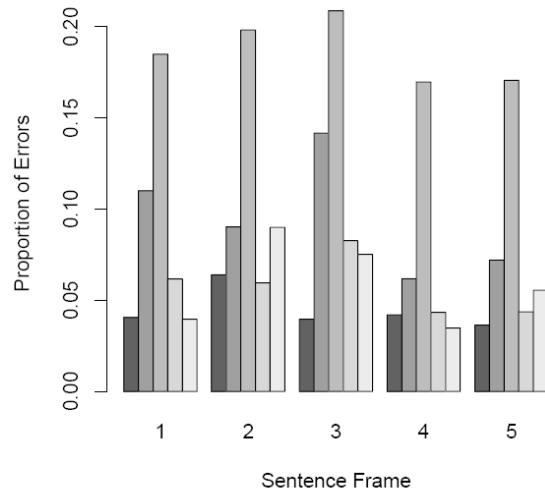


Figure 4: Means for interaction of synthesizer and sentence frame for phone-level edit distance. The shaded bars in each grouping represent the means for each synthesizer (A – E in alphabetical order) within each sentence frame.

analysis (LOA) was a significant main effect and participated in significant interactions with both synthesizer (SYN) and sentence frame (FRM).

Intelligibility by Level of Analysis

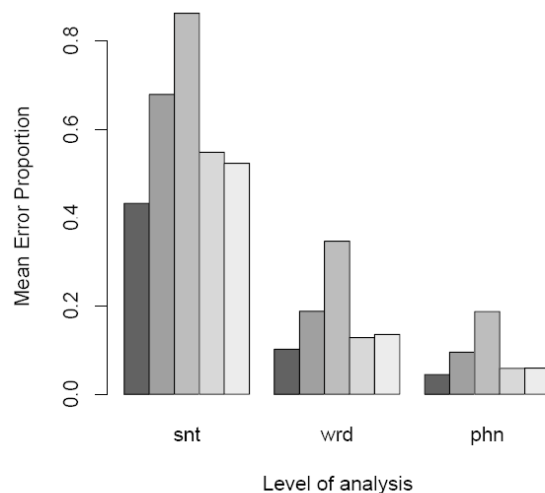


Figure 5: Comparison of mean proportion of errors for each synthesizer (shaded bars) at each level of analysis.

Table 2: Summary table from multi-level ANOVA. All terms are $p < .001$.

Term	Df	SSQ	MSQ	F
LOA	2	187.1	93.6	2714.6
SYN	4	34.4	10.9	121.8
FRM	4	7.7	1.9	29.4
LOAxSYN	8	6.8	0.9	30.9
LOAxFRM	8	2.7	0.3	17.4
SYNxFRM	16	5.2	0.3	4.1
LOAxSYNxFRM	32	2.1	0.1	3.3

3.5. Hand correction

To determine the consequences of carefully hand-correcting the phonetic transcriptions of listener response data, phonetic edit distances computed from hand-corrected versus uncorrected dictionaries were compared in an analysis of variance using synthesis system, sentence frame, and correction as factors in a 5 x 5 x 2 design. As expected from all the previous analyses, this analysis revealed significant main effects of synthesizer and sentence frame as well as a significant interaction of sentence frame and synthesizer. There was also a significant main effect of correction, with carefully corrected transcriptions having overall lower edit distances than did uncorrected data ($F[1,29]=41.72, p < .001$). Crucially, correction did not interact with any other factor. Thus, while hand correction of phonetic transcriptions for listener responses did result in lower edit distances overall, it had no further consequences for interpreting the results.

3.6. Study size

Study size can be varied either by changing the number of listeners involved, or by changing the number of stimuli per condition that are presented to each listener. In the latter case, reducing the number of stimuli per condition can reduce the total amount of time required to run a study, or allow more conditions to be explored with the same total number of stimuli. Reducing the number of listeners can also reduce the amount of time required to run a study (reducing cost if listeners are paid), or if listeners are grouped in different conditions, allow more conditions to be explored with the same total number of subjects.

We simulated the relative costs of reducing the number of subjects per condition versus reducing the number of stimuli by repeatedly randomly discarding 50% of the subjects or sentences in a balanced manner, and recalculating the results based on the randomly selected subset. Results from 50 such simulated smaller experiments are presented in Figure 6, which shows boxplots for the results when subjects are randomly discarded (left panel) and when sentences are randomly discarded (right panel). Each boxplot represents the median error proportion (horizontal line in each box), the interquartile range (box vertical extent), and the full range of results (whiskers) sans data points identified as outliers (circles). The amount of variability (as indicated by

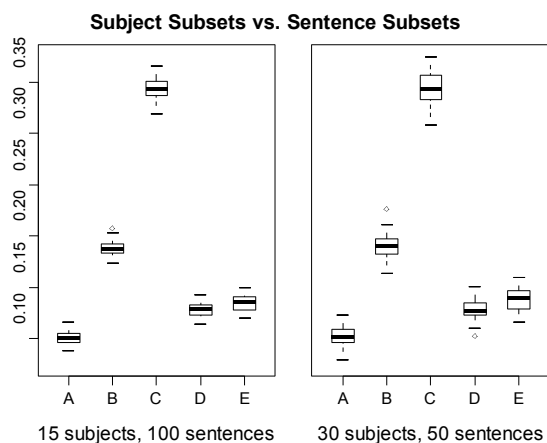


Figure 6: Comparison of simulations using a reduced number of subjects (left panel) versus a reduced number of stimuli (right panel).

interquartile range) is clearly larger for simulations based on

discarding sentences than for simulations based on discarding subjects.

Given that it is less costly—in terms of experimental power—to reduce listeners than to reduce sentences, we next simulated a series of studies of size ranging from one listener per group (total $N=5$) to 6 listeners per group (total $N=30$, i.e., the original study). In this simulation, we sought to determine how many listeners were needed to retain significant pair-wise differences between synthesis systems. The results of this analysis are shown in Figure 7 where each small panel presents the average t-values for 50 comparisons between one pair of TTS systems with various numbers of subjects. Red dotted lines indicate the nominally significant level ($p < .05$) of t (without correction for multiple tests). Error bars indicate to total range of t-values observed. Significant differences between systems D and E were never observed. Significant differences between system A and systems D and E are sometimes lost with even a reduction from 30 to 25 subjects. Virtually all other pair-wise comparisons remained significant with only 5 or 10 listeners (i.e., one or two listeners per group).

4. Discussion

A variety of different analyses were presented to examine intelligibility measures at different levels of analysis. At the sentence level, the absolute overall ranking of the five TTS systems differed slightly from the other two levels. System E had a lower proportion of errors than System D in the sentence-level analysis, but a higher proportion in other analyses. However, the differences between these two systems were extremely small and not statistically significant in any analysis. Hence, both systems should really be considered to share a single rank in all analyses. With this qualification in mind, it seems safe to conclude that for merely ranking the intelligibility of TTS systems, it makes little difference whether one uses a simple “sentences correct” measure or a more labor-intensive phonetic edit distance measure.

It is often important, however, to be able characterize how much more intelligible one system is compared to others. For instance, in selecting a TTS system for a specific application, one may want to consider multiple factors including intelligibility, naturalness, preference for a specific voice gender, etc. In weighing these factors to arrive at a final decision, knowing the amount of intelligibility difference between two systems is essential. That is, one may be willing to accept a small, but not a large, loss of intelligibility in favor of a more natural or pleasing voice. Our results here suggest that sentence-level scoring may obscure the magnitude of the differences between systems. While the present analysis was sufficiently well powered to detect the overall differences between most of the TTS systems at all levels of analysis, there is concern that screening studies run with fewer listeners and/or a smaller number of utterances per synthesizer would fail to detect differences with sentence-level scoring that would be detectable with word- or phone-level scoring.

It is, of course, clear that developers of TTS systems need analyses at the phonetic level to diagnose specific strengths and weaknesses with systems. In that case, an encouraging finding from the present study is the absence of secondary effects of carefully hand-correcting phonetic data. Although our efforts to carefully interpret the phonetic intent of the subjects did result in overall lower error rates for all systems,

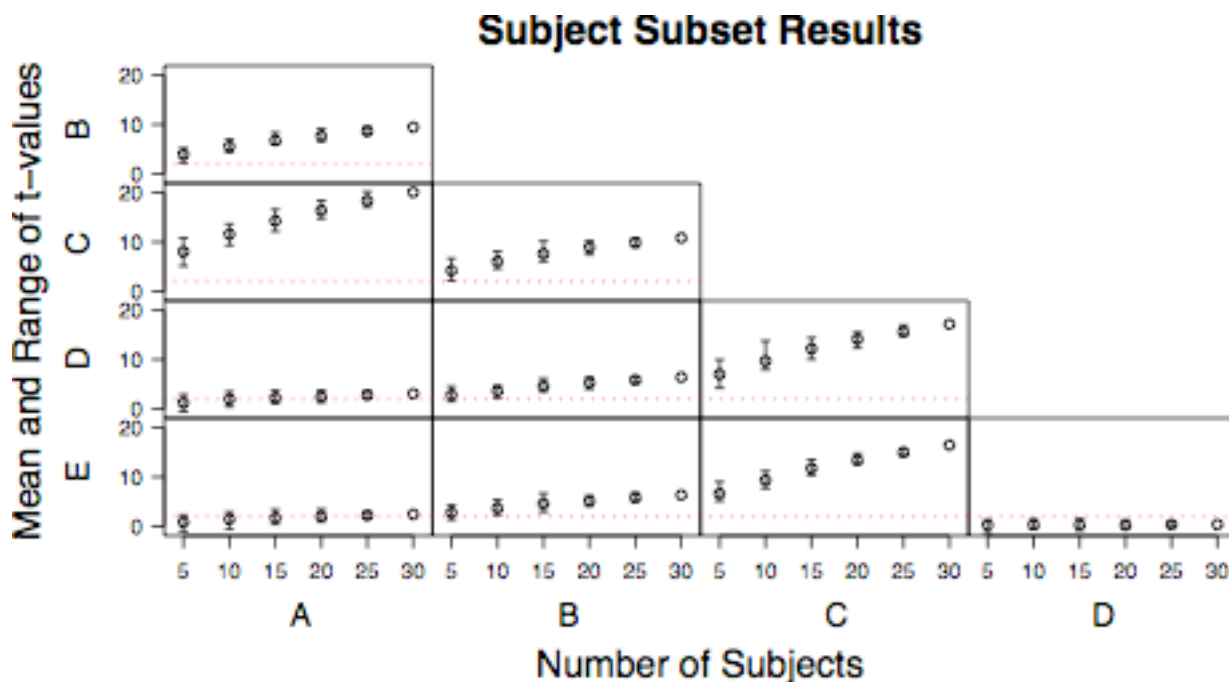


Figure 7: Average t-values for all pairwise comparisons among the five synthesizers as a function of the number of subjects. Error bars reflect total range of t-values obtained in 50 simulation trials. See text for explanation.

there is no evidence that these efforts would have consequences for conclusions drawn from the study. This in turn suggests that it should be possible to develop relatively automated scoring procedures based on dictionaries that include entries for highly probable typos, misspellings, and the like.

Another advantage to using phonetic edit distances is that the edit distance data can be further analyzed to diagnose specific phonetic strengths and weaknesses within a system, or to discover differences between systems with indistinguishable gross intelligibility scores. For example, while systems D and E in the present analyses have indistinguishable total edit distance scores, we found that responses to system E had a greater number of deletions, while responses to system D had greater numbers of insertions and substitutions. These differences in the types of errors listeners make on one system versus another may prove to be of diagnostic value. It is also a simple matter, during the computation of phonetic edit distances, to tabulate a confusion matrix of stimulus phones versus response phones, allowing systems to be examined and compared by phonetic classes. For example, responses to system E were more likely to delete voiced phones or replace them with voiceless ones, while responses to system D were more likely to substitute labials with nonlabials.

Finally, simulations of smaller experiments with different numbers of listeners and stimuli highlighted the importance of using a large number of stimuli per synthesizer, relative to the number of listeners. It is interesting to speculate that this may be particularly true of studies using concatenative TTS systems because of the very large number of unique concatenation sequences such systems may employ.

5. Conclusions

If one is only interested in ranking the relative intelligibility of several TTS systems, sentence-level scoring of SUS response data may be adequate, but it may mask the magnitude of real

differences between systems. For more detailed analyses, phonetic edit distance is a more attractive measure. While the amount of effort needed to obtain phonetic-level edit distances is greater than that needed for a words or sentences correct measure, we found that little is gained by investing large amounts of effort in screening and interpreting listener responses. Instead, a more automated (and probably more objective) approach yields slightly higher overall error rates, but does not otherwise appear to influence conclusions one might draw from the data.

6. Acknowledgements

Work supported by NIDCD grant # R42DC006193 and Nemours Biomedical Research.

7. References

1. Nye, P.W. and J.H. Gaitenby, The intelligibility of synthetic monosyllabic words in short, syntactically normal sentences. *Haskins Laboratory Status Reports on Speech Research*, 1974. **37/38**: p. 169-190.
2. Benoit, C., M. Grice, and V. Hazan, The SUS test: A method for the assessment of text-to-speech synthesis intelligibility using Semantically Unpredictable Sentences. *Speech Communication*, 1996. **18**(4): p. 381-392.
3. Benoit, C., An Intelligibility Test Using Semantically Unpredictable Sentences - Towards the Quantification of Linguistic Complexity. *Speech Communication*, 1990. **9**(4): p. 293-304.
4. Bunnell, H.T., et al., Automatic personal synthetic voice construction. *Proceedings of InterSpeech 2005*, Lisbon, Portugal, 2005.