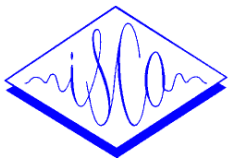


First International Conference on
Spoken Language Processing (ICSLP 90)
Kobe, Japan
November 18-22, 1990

EVALUATING SYNTHESISER PERFORMANCE: IS SEGMENTAL INTELLIGIBILITY ENOUGH?

Kim Silverman, Sara Basson, Suzi Levas

NYNEX Science and Technology
500 Westchester Avenue, White Plains, NY 10604, USA

ABSTRACT

Laboratory-based evaluations of synthetic speech often suggest that it is as intelligible as natural speech. Yet studies and experience in applied settings do not confirm this. We identify two issues underlying this discrepancy: (i) neither the speech material nor the listeners' task in typical evaluations sufficiently represent application conditions, and (ii) tests should measure the cognitive load accompanying a given intelligibility score. Our preliminary data for two commercial synthesisers underline both issues: the two synthesisers were ranked in one order on a segmental intelligibility test, but in the opposite order on a more application-like comprehension test. For the latter test, listeners' accuracy may be related to their performance speed. Results are related to the underlying issues, and two approaches to measurement of cognitive load are suggested.

LABORATORY TESTS VERSUS APPLIED SETTINGS

There is an interesting paradox concerning the quality and usefulness of synthetic speech technology. On the one hand, laboratory tests and vendors' claims indicate that the quality is quite high. For example, Greene, Logan and Pisoni [2] found that the best synthetic speech could be perceived by listeners with better than 96% accuracy. When natural speech tokens are used in the same tests, intelligibility scores are between 96% and 99%. Data such as these suggest that synthetic speech is achieving a degree of intelligibility comparable to natural speech.

On the other hand, those relatively few studies of synthetic speech performance in more applied settings yield a somewhat different picture. The speech sounds unpleasant, unnaturally monotonous, and boring. It is very difficult to follow the meaning of synthesised passages. To do so requires more concentration than for natural speech; the effort is fatiguing and the task is annoying. Even isolated short words, when spoken by a synthesiser, interfere with memory more than when spoken by a human (for a more detailed review and discussion see [4] and [7]).

This, then, is the paradox. If, in laboratory tests, synthetic speech proves to be about as intelligible as natural speech, then why — unlike natural speech — is it so difficult to understand in applied settings? We believe there are two related issues underlying this apparent contradiction.

Issue I: The Material and the Task

The first issue concerns the nature and purpose of the spoken material, in laboratory tests and in real applications. Most intelligibility tests require listeners to identify single, short, isolated words (or even nonsense syllables). In such tests a synthesiser is scored according to the proportion of correctly-identified segments. Most applications, by contrast, employ more extended material: connected speech rather than isolated words. And in applied settings the purpose of the speech is to convey information: listeners need to understand the meaning of the connected speech, and generally take some action on the basis of this meaning. Consequently it is unclear how to predict synthesiser performance in applied settings on the basis of segmental identification scores alone. To do so requires an underlying assumption that we can account for the perception of speech in terms of the intelligibility of the segmental structure in isolated words.

This assumption almost completely excludes the perceptual importance of structure at the suprasegmental level. The prosodic organisation imposed by synthesisers on words when they are strung together into connected sentences cannot be properly tested by evaluations based on words spoken in isolation. Yet there is pervasive and persuasive evidence in the literature that sentence-level suprasegmental structure makes crucial contributions to many levels of speech perception, including pragmatic interpretation, structural disambiguation, reference resolution, pruning inferences, lexical access, and even facilitating phoneme identification (see, e.g., [1], [6], [7], [9]).

To be fair, tests of segmental intelligibility have the advantage of being diagnostic: they can isolate incorrect phonemes and pinpoint patterns of confusion. Nevertheless, in an effort to reduce the gap between predicted and actual performance in applied settings, we are investigating the use of sentence-length material rather than words or syllables presented in isolation. In the test we describe in this paper, the listeners' task is not to transcribe words, but to answer questions designed to assess whether they have understood the content of the sentences. In this sense we believe the task is closer to a real application than is segmental transcription. In addition, the topics of our sentences are taken from real or potential applications.

Issue II: Cognitive Load

Most applications require listeners to respond to the meaning of the speech, and typically listeners are at the same time performing some other task involving the hands, eyes, and some degree of attention. This task may be seemingly simple, such as concurrently writing down a synthesised address or telephone number. Alternatively it may be a more obviously demanding process, such as air traffic control or landing an aircraft. The important point is that in applied settings listeners are typically less able to concentrate solely and fully on the speech than they can in laboratory tests. This is the second issue which we believe clouds the evaluation of synthesiser performance: we propose that any assessment of a synthesiser needs to take into account the load that that particular synthesiser's speech places on the listener's necessarily limited cognitive resources.

There are two reasons why we believe cognitive load to be important. Firstly, if a synthesiser demands very much concentration and effort to understand, then this will reduce

both (i) the amount of attention that a listener can give to other concurrent tasks, and (ii) the amount of time before performance is affected by fatigue. Consequently a synthesiser that imposes a higher cognitive load to achieve an acceptible level of intelligibility will interfere more with people's performance (speed, accuracy, fatigue) on their concurrent tasks and/or on their responses to the synthesised messages. Such cognitive load characteristics of component tasks are important considerations in the design of human-computer interfaces (e.g. [5]).

The second reason why cognitive load is important is more fundamental: without this information the results of any intelligibility test are difficult to interpret and may even be misleading. We will illustrate this problem with a hypothetical example [1]. Suppose fifty listeners participate in a comprehension test on the output of synthesiser X, and another fifty take the same test with natural speech. The ranges of scores for the two voices overlap, but the means differ in the direction of better comprehension of the synthetic speech than of natural speech. The histograms of the hypothetical individual scores are shown in Figure 1.

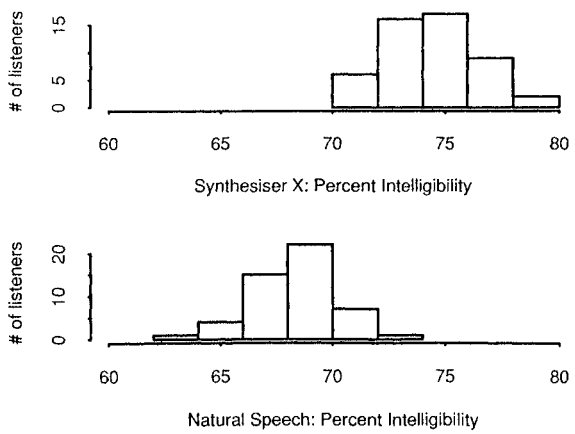


Figure 1: Hypothetical Intelligibility Scores

At first glance, such a result might tempt one to conclude that at least for this particular type of test, synthesiser X is more intelligible than natural speech. Let us call this conclusion **interpretation I.**

A simple model of the listener would be that the score achieved on this test is a linear function of the quality of the synthetic (or natural) speech, but also of that listener's individual ability (a combination of innate aptitude and prior experience), the amount of effort that listener brings to bear during the test, and some sampling error:

(1) $\text{score}_{ij} = a.\text{sq}_j + b.\text{ia}_i + c.\text{effort}_{ij} + \text{error}_{ij}$

where i = listener
j = voice (synthetic or natural)
sq = speech quality
ia = individual ability
effort = overall effort during the test
error = measurement error

[1] The means and direction of the difference in this simulation are taken from the real data presented in Table 3 of [3]. Specifically, they are from the second half of a comprehension test, where after some practice listeners achieved higher scores with synthetic speech than with natural speech. The current discussion is not a criticism of those results per se, but rather merely uses them to raise a more general problem.

The first variable (sq) is constant, the other three we model as gaussian random variables. It is the second of these, **effort**, that we consider here. We understand greater effort to mean that the listener is concentrating harder, i.e. devoting more of his/her cognitive resources (attention, memory, inferencing capabilities, processing time, etc) to the task. Figure 2 plots, on the same axis as Figure 1, the relationship between test score and cognitive resources used by fifty simulated listeners for each voice. The lines through each cloud are least-squares regressions. They have a positive slope because as listeners try harder, they perform better (as expressed in equation **1**). Points above the line in each cloud represent individuals with greater individual ability, points below represent those with less. Each regression accounts for about 33% of the variance, as we would expect from equation **1**.

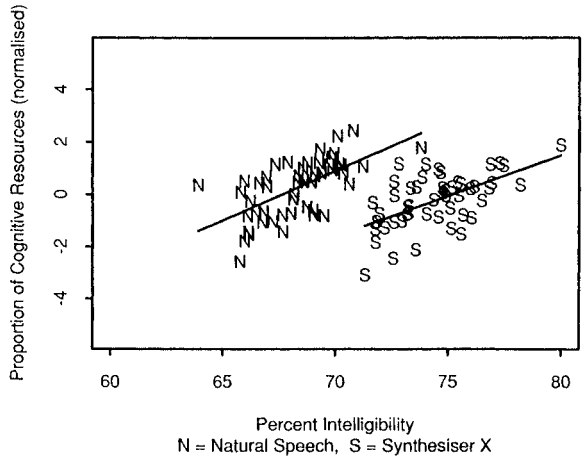


Figure 2: Intelligibility vs Cognitive Load (**Interpretation I**)

One thing that Figure 2 sheds light on is the region of overlap between the intelligibility scores of synthesiser X and natural speech (i.e. between about 71% and 74%). Specifically, it shows that in order to achieve a test score in this region, listeners need to concentrate much harder if they hear natural speech than if they hear the synthetic speech.

Figure 2 underlies **interpretation I**, and shows how it assumes that (i) the two groups of listeners use similar amounts of effort, and therefore (ii) the results shown in Figure 1 arise because the synthesiser is **more intelligible** than natural speech.

There is, however, an alternative way in which the differences in Figure 1 may have arisen. Suppose the synthesised speech sounded so unnatural and difficult to follow that it induced listeners to concentrate much harder than they are accustomed when listening to natural speech. Let us call this **interpretation II**. The relationship between effort and performance would then be as shown in Figure 3 (in the next column).

Figure 3 (and **interpretation II**) tell a totally different story: they indicate that a score in the overlap region will demand a far greater proportion of a listener's cognitive resources if the voice is synthetic than if the voice is natural. Moreover, extrapolating the regression lines shows that underlying the test results achieved by synthesiser X is a greater cognitive load than we would predict on the basis of the effort/performance relationship for natural speech. Therefore the results in Figure 1 arise because listeners concentrated disproportionately harder for synthetic speech.

The higher test scores for the synthesiser hide the fact that despite the direction of the difference, synthesiser X is in reality **less intelligible** than natural speech.

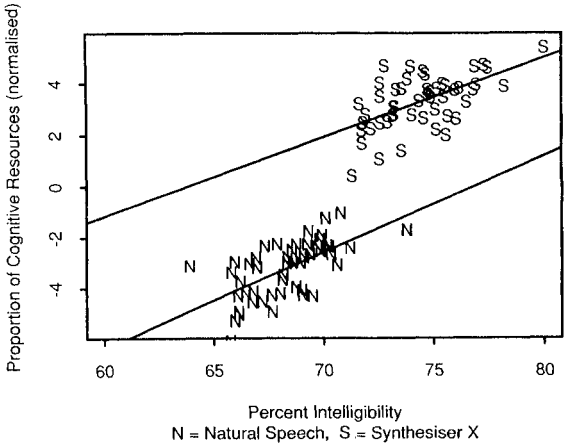


Figure 3: Intelligibility vs Cognitive Load (**Interpretation II**)

To summarise, there is a discrepancy between the results of laboratory-based tests of synthesiser intelligibility and the reported performance of the technology in more applied settings. One reason is that neither the spoken material nor the listeners' tasks are directly comparable. More specifically, it is not clear how accuracy in transcription of short, isolated words relates to understanding the meaning of connected speech in order to achieve some task-oriented goal. The second reason is that traditional evaluations do not take into account the demands synthetic speech places on a listener's cognitive resources.

Our current aims are (i) to develop both speech material and a more global task that more closely approximate applied settings, (ii) to compare the results of this task with a test of segmental intelligibility, and (iii) to begin investigating measurements of cognitive load that are relevant to perception of synthetic speech.

Method

For the initial phase of these experiments, two commercially available formant synthesisers were selected for assessment and comparison, referred to henceforth as synthesisers "A" and "B".

The segmental intelligibility test developed at Bellcore [6] was used to assess each synthesiser. Its 312 test items were designed to rigorously measure consonant intelligibility in initial and final position, using nonwords as well as words, and including all permissable consonant clusters. For each stimulus item, transcriptional errors on particular target segments were identified and tallied.

A more global comprehension test has been developed internally. The 88 test sentences were designed to simulate actual applications where speech synthesis may be used; for example, accessing database information or providing travel instructions. Sentences in the set are on average 20 words long, and vary in difficulty, complexity, and length. Each test item consists of one of these sentences spoken by the synthesiser under test, followed by a question asked by a recorded adult male human. Each question can be answered with "yes," "no," or "can't tell from the information provided." The faster that listeners answer questions, the more items they hear. Sample test items are:

Synth: Model numbers 3871, 3872, and 6528 are unavailable.
Human: *Are all models available?*

(Correct answer: NO)

Synth: If you need additional information, press 1; to make a selection, press 2; if you'd like to speak to an operator, press 3.
Human: *Will #4 disconnect the call?*

(Correct answer: CAN'T TELL)

A concurrent visual-motor task was included in order to increase test sensitivity and more closely model an application environment. While answering the comprehension questions, listeners track a randomly moving square on a computer screen by moving the mouse.

80 listeners so far have participated, with about 20 for each synthesiser in each test. They were encouraged with financial incentives to perform at the peak of their ability for both the speech and (in the comprehension test) the tracking task.

In each of the tests, the synthesisers spoke at their own default speaking rate. We assume that these are determined by the manufacturers to yield maximal clarity of the speech. Synthesiser B spoke somewhat slower than A (ie at 93% of the speed); we shall return to this point in the discussion.

Results

We report here briefly on the data so far collected. According to the segmental intelligibility test results, synthesiser A has a significantly higher phoneme intelligibility than B (77% vs 70%, $t_{43} = 4.40$, $p<0.0001$). The raw data are summarised in Figure 4.

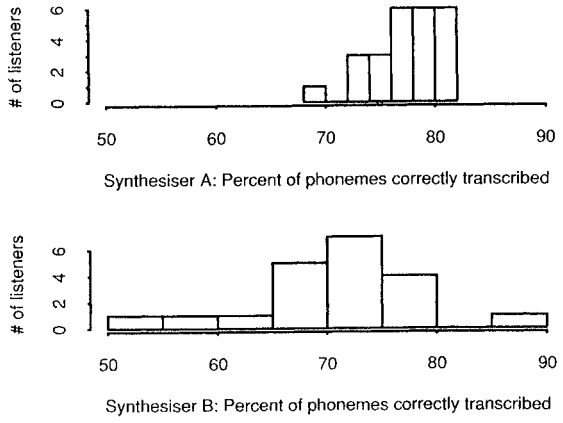
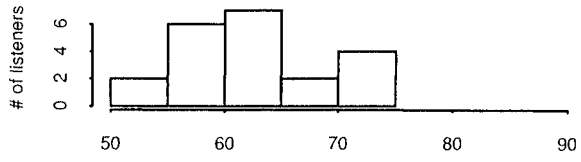


Figure 4: Segmental Intelligibility Test Results

In the comprehension test, there was somewhat more overlap between the synthesisers. More interestingly, the two synthesisers ranked in the opposite order in this test: listeners answered a higher proportion of items correctly for synthesiser B than for A (69% vs 63%, $t_{37} = 2.242$, $p<0.05$) The raw data are summarised in Figure 5.

But these accuracy scores tell only part of the story: listeners heard on average more items from synthesiser A than from B (77 vs 69, $t_{37} = 14.684$, $p < 0001$) in the 20-minute test-session. These distributions show almost no overlap at all, as can be seen in Figure 6.



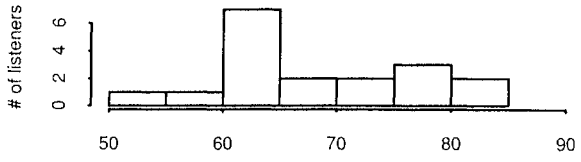
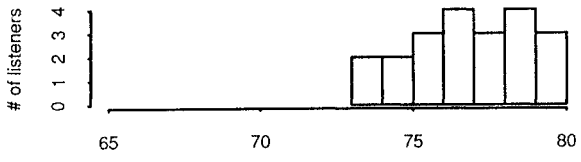


Figure 5: Comprehension Test Results



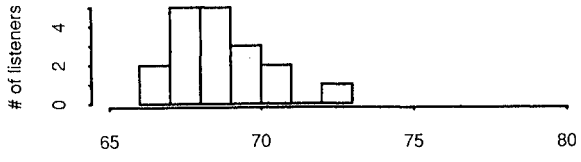


Figure 6: Total number of Comprehension Items Answered

Discussion

These preliminary results suggest two interesting reversals. The first is that although listeners more accurately perceived the phonemes in the speech of synthesiser A, they more accurately comprehended the speech of B. This underlines the first problem raised in the introduction: it is not clear how accuracy in transcription of short isolated words relates to the overall process of understanding speech. One thing we notice is that because the segmental intelligibility test used isolated monosyllables to assess word-initial and word-final consonants, it is really focussing on consonants that are utterance-inital and utterance-final. These are by far the minority in sentence-length material.

The second reversal is within the comprehension task: although listeners had a higher percentage accuracy with B, they answered more items for synthesiser A. This leaves us with two possible (though not mutually-exclusive) interpretations. One is that since synthesiser B spoke at 93% of the rate of A, it allowed listeners more time to identify and disambiguate the speech, which consequently increased their accuracy — although at the expense of their performance speed. The other possibility is that listeners who heard B had more time to decode the speech than they needed, consequently (i) at the default speaking rate listeners who hear B should be more able to attend to a concurrent task than listeners who hear A, and (ii) if B's speech had been faster (allowing them to hear more items) they would have responded just as accurately.

In general, we believe that speed/accuracy trade-offs provide one way to assess the overall mental processing demands of synthetic speech, and we intend to refine our techniques for measuring these while controlling for speech rate We also shall look to the relationship between speech comprehension and performance on the concurrent task, as another way to estimate the cognitive load associated with understanding the synthetic speech. We hope to report our first results from the tracking task in the near future.

Acknowledgements

We are grateful to Julie Silverman for implementing and carrying out the Bellcore test, to Erik Urdang and Ashok Kalyanswamy for working magic with C and AWK, to Sheri Waltzman for running the comprehension and tracking tasks, and to Judy Spitz and Murray Spiegel for their encouragement and critical insights. Errors and omissions, nonetheless, remain our own responsibility.

References

[1] Cutler, A. and Darwin, C.J. (1981) Phoneme-monitoring reaction time and preceding prosody: Effects of stop closure duration and of fundamental frequency. *Perception and Psychophysics*, **29**, 217-224.

[2] Greene, B.G; Logan, J.S. and Pisoni, D.B. (1986) Perception of synthetic speech produced by rule: Intelligibility of eight text-to-speech systems. *Behavior Research Methods, Instruments and Computers*, **18**, 100-107.

[3] Pisoni, D.B, Nusbaum, H.C. and Greene, Beth G. (1985) Perception of synthetic speech generated by rule. *Proceedings of the IEEE*, **73**, 11, 1665-1676.

[4] Pisoni, D.B., Manous, L.M. and Dedina, M.J. (1987) Comprehension of natural and synthetic speech: Effects of predictability on the verification of sentences controlled for intelligibility. *Computer Speech and Language*, **2**, 303-320.

[5] Schneiderman, B. (1987) *Designing the User Interface: Strategies for Effective Human-Computer Interaction.* Addison-Wesley.

[6] Silverman, K.E.A. (1986) F0 cues to voicing depend on intonation: the case of the rise after voiced stops. *Phonetica (Special Issue on Prosodic Cues to Segments)*, **43**, 76-91.

[7] Silverman, K.E.A (1987). *The Structure and Processing of Fundamental Frequency Contours* PhD Dissertation, Cambridge University.

[8] Spiegel, M., Altom, M.J., Macchi, M and Wallace, K. (1988) Using a monosyllabic test corpus to evaluate the intelligibility of synthesized and natural speech. *Proceedings of the American Voice I/O Systems Conference.* (Extended Paper available from the Authors).

[9] Wales, R.W. and Toner, H. (1979) In W.E. Cooper and E.C. Walker (eds) *Sentence Processing: Psycholinguistic Studies Presented to Merril Garrett.* Halstead Press.