ASSESSING THE INTELLIGIBILITY AND PROCESSING SPEED OF PROCESSED SPEECH

KERRIE MACKIE PHILLIP DERMODY RICHARD KATSCH

Speech Communication Research Section
National Acoustic Laboratories
126 Greville, St., Chatswood. N.S.W. 2067. AUSTRALIA.

ABSTRACT - The present study examines evaluation measures designed to assess the intelligibility, and speed of processing of natural speech with harmonic distortion. The results indicate that even for highly intelligible processed natural speech delays in processing time are a consequence of poor acoustic phonetic information. The results also indicate the value of including more sensitive tests of speech intelligibility in evaluation protocols for speech transmission evaluation.

INTRODUCTION

In the development of voice communication devices, listening tests have always been an important part of evaluation procedures. The primary focus of these evaluation procedures has been the assessment of the intelligibility and quality of the speech through the device compared to some arbitrary standard. However, with continuing improvements in speech transmission systems the aim of the assessment procedure has changed to one which compares the output of speech transmission systems to listening results for natural speech.

The improvement in the quality and intelligibility of modern communications systems has produced a need for more sensitive assessment procedures which are appropriate to the evaluation of devices such as hearing aids where intelligibility is typically high, to the evaluation of synthetic speech produced by text to speech systems where intelligibility has a wide range of adequacy. The speech assessment of hearing aids is an active area of work and the evaluation of synthetic speech has produced some improved measures for speech evaluation(1).

In the present study we investigated measures of intelligibility and of processing speed to determine their relationship. Pisoni, et al (1) have applied a range of intelligibility and processing measures to the assessment of synthetic speech. They report that listeners have slower response times to synthetic speech compared to natural speech and they concluded that the increase in processing time is due to the poorer segmental intelligibility of the synthetic speech stimuli compared to natural speech. That is, difficulties at the acoustic phonetic level for synthetic speech underlie later processing time increases. In the present study we investigate measures of intelligibility and processing speed to determine their relationship for natural speech stimuli which have harmonic distortion.

EXPERIMENT 1

The first experiment was designed to demonstrate that the harmonic distortion of the stimuli increased the processing time for listeners in a manner similar to that reported for synthetic speech (1). The processing speed task chosen was the auditory lexical decision task. In this task the listener hears a stimulus and must decide as quickly as possible whether it is a word or a non word. Stimuli consisted of monosyllabic English words or pronounceable non-words. Pisoni, et al (1) reported that the lexical decision task showed slower reaction times for synthetic speech relative to natural speech, although for both synthetic and natural speech the relationship between words and non-word reaction times was similar. Pisoni, et al (1) concluded that synthetic speech is processed in a similar manner to natural speech at the lexical level, but that the impoverished acoustic-phonetic structure of synthetic speech led to its longer processing time overall.

STIMULI

The speech stimuli were recorded by an adult male speaker onto a computer speech storage/editing system using a 12 bit analogue to digital converter at a sampling rate of 36K samples per second. The stimuli were copied into three separate disc files which were then separately processed using an algorithm based on Schroeder (2) in which noise is added to the digitally sampled speech randomly over a specified time to produce harmonic distortion of the original waveform. The speech produced can be expressed as a change in signal-to-noise (S/N) ratio compared to natural speech. The S/N ratio is determined by the amount of distortion which is added per sample. In each file the speech was processed to give a S/N ratio of either 0, +3, or +6dB. The speech in each file was highly intelligible. This is attributable to the high sampling rate of the speech and the fact that the distortion technique is based on random noise addition per sample. When lower sampling rates are used, the result is considerably more degradation of the sampled speech for the same signal to noise ratio (2).

The listening tests were carried out on subjects seated in an audiometric test booth and speech was presented to them binaurally via headphones (type TDH49). The computer randomly selected the speech stimuli from the disc files and presented the stimuli, recorded the subject's responses and the reaction time in milliseconds for each stimulus.

RESULTS

Figure 1 presents the results for the lexical decision task and shows that in all conditions listeners respond faster to words than non words. The difference between the processed speech and the natural speech are reflected in reaction time differences only. That is, the lexical decision task reveals processing time differences between the different speech conditions and natural speech but not qualitative differences in processing the words and non-words. These results are similar to the results of Pisoni, et al (1) for synthetic speech compared to natural speech.

EXPERIMENT 2

In order to explore the relationship between intelligibility and processing speed we investigated intelligibility and other processing speed tasks. In these tasks the stimuli used were limited to a closed set of six highly confusable CV syllables (stop consonants plus the vowel /a/). This limited set was chosen because it provided a demanding discrimination task for listeners and would therefore be sensitive to both intelligibility and processing time differences. These stimuli were processed to give three levels of distortion, as in experiment 1.

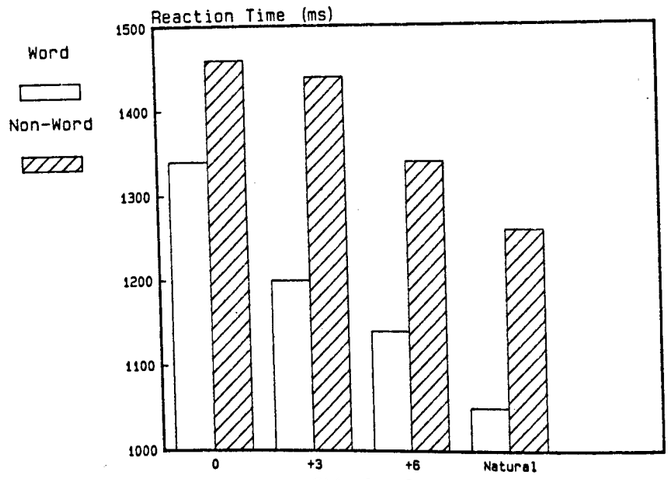

Figure 1
Results of the lexical decision task for processed and natural speech.

To establish that the stimuli in each of the conditions were all highly intelligible subjects were presented with the stimuli from the three S/N conditions in a randomised order at 60 dB SPL. There were ten presentations of each stimulus and the subject's task was to press one of six buttons corresponding to the speech sound they heard. Response time as well as accuracy were recorded. The accuracy results showed that the speech stimuli in each S/N condition were intelligible with all scores approaching 100 percent (99.3%;99.7% & 98.2% for the 0,3 & 6 dB S/N conditions respectively). The reaction time data showed slight differences between the conditions with the slowest reaction time for the most distorted speech (the 0 dB S/N condition) lending some support to the claim by Pisoni,et al (1) that measures of processing speed can often show differences between speech signals which are not differentiated on suprathreshold intelligibility tests because of ceiling effects. However, the effects here are very slight and do not conclusively distinguish the test conditions.

These same stimuli were also used in two processing speed tasks including a two alternative forced choice task (2AFC) and a forced choice comparison task.

In the 2AFC task the subject heard a stimulus and chose from two response alternatives which stimulus was presented. The emphasis in the task is on speed of response although accuracy as well as reaction time are recorded. The results indicate a small but significant difference between the three S/N conditions, showing an increase in processing time with an increase in

the amount of distortion. Accuracy results again indicate high intelligibility and no differences between the different conditions.

In the forced choice comparison task, the subject heard an undistorted CV syllable followed one second later by a tone and another CV syllable which was distorted. The subject's task was to make a yes/no judgement as to whether the second syllable was the same as the first syllable. Subjects were required to make these judgements as quickly as possible. The results indicated that for all conditions there was no difference in accuracy judgements but that subjects required longer processing times with the more distorted stimuli.

These data from the processing measures show consistent processing speed differences between the stimulus conditions in the absence of intelligibility differences, suggesting that the increased processing time is the result of higher level cognitive factors rather than difficulties at the representational level,as found in synthetic speech by Pisoni, et al (1).

To investigate this discrepancy we carried out two further intelligibility measures to ensure that the segmental intelligibility of the stimuli in each of the conditions was in fact the same. These measures included an adaptive speech test using the PEST procedure (3) and a stimulus repetition task (4).

In the PEST procedure the subject was required to press one of two buttons in front of them to indicate which stimulus was presented. The response alternatives changed on each trial and were displayed on a screen. The subject's responses were monitored for proportion correct and if this fell below a specified criterion, then the stimulus level was increased. If it fell above the specified criterion then the stimulus level was lowered. In this way, the presentation level of each stimulus was changed depending on the performance of the subject. The testing was continued until a specified criterion of performance was achieved. The results for the PEST procedure are expressed as the dB level at which the speech recognition threshold was achieved. The results show significant differences between the thresholds for the three conditions. The recognition threshold for the least distorted conditon (+6) is 2 dB better than for the +3 condition, which in turn is about 2dB better than the 0 condition.

The second intelligibility measure contrasted the subject's performance when the stimulus was presented once per trial compared to three repetitions of the stimulus before a response was required. Clark, Dermody & Palethorpe (4) found this procedure differentiated between synthetic and natural speech, with natural speech showing a significant increase in intelligibility with three repetitions while synthetic speech did not improve. In the present study the speech stimuli were presented near the 50% correct level (based on the 0dB condition). Subjects were presented with the single repetition or the three repetition condition in a counterbalanced design. The results indicate that there is a repetition effect in each test condition and that the least distorted condition produces the greatest repetition effect. That is, speech in each of the test conditions is processed in a similar way to natural speech in the Clark,et al (4) study, but with a slightly reduced effect because the speech stimuli were more distorted. These results are similar to the results for the lexical decision task which also showed that the distorted stimuli behaved in a similar manner to natural speech in that case with longer processing time.

CONCLUSIONS

The results from experiment 2 suggest that even when processed natural speech is highly intelligible at suprathreshold levels, it can still produce slow processing times compared to natural speech. The results of a sensitive speech intelligibility task using the PEST procedure indicates that there are slight but significant differences for recognition of the processed speech which produce the slower processing times. This result is consistent with the notion that poorer acoustic phonetic processing will slow processing time for synthetic speech which is impoverished compared to natural speech as suggested by Pisoni, et al (1). The present study extends this finding to natural speech that has had noise added.

The present results suggest i) that high intelligibility at suprathreshold levels should not be used as a sole criterion for speech transmission if comparision with natural speech is intended and ii) that sensitive measures of both intelligibility and processing time can be used to differentiate processed natural speech from natural speech in listening performance when suprathreshold intelligibility of the processed speech is equivalent to natural speech.

REFERENCES

(1) PISONI,D.,NUSBAUM,H.& GREENE,B. (1985)"Perception of synthetic speech generated by rule",Proceedings of IEEE, 73,1665-1676.

(2) SCHROEDER,M.(1968)."Reference signal for signal quality studies",Journal of Acoustical Society of America, 44,1735-1736.

(3) TAYLOR,M.& CREELMAN,C.(1967)"PEST: efficient estimates on probability functions",Journal of Acoustical Society of America,41,782-787. for the three distortion

(4) CLARK,J.,DERMODY,P.& PALETHORPE,S.(1985)"Cue enhancement by stimulus repetition: natural and synthetic speech comparisons",Journal of Acoustical Society of America, 78, 458-462.