Speech Quality and the Gating Paradigm

S.G. Nooteboom and G.J.N. Doodeman
Eindhoven, the Netherlands

1. Introduction

The correct recognition of a word from speech can often take place when only part of the spoken word form has been heard (e.g. Marslen-Wilson, 1980). In such a case the remainder of the word is, in some sense, redundant information. Of course, redundant does not mean superfluous. Redundancy serves to make speech communication less vulnerable to all kinds of degradation of the 'ideal' speech signal, caused for example by sloppy articulation, external distortion, or a hearing deficit in the listener.

In the present experiment we have set out to measure the effect of differences in speech quality, caused by differences in degree of data reduction in LPC vocoder speech, on the relative number of speech sounds needed for correct recognition of polysyllabic words. For this purpose we used an adaptation of the 'gating paradigm' introduced by Grosjean (1980).

Our aim was twofold. We wanted to find out whether we could obtain a reliable and relatively easily applied measure of speech quality. We also wanted to see whether the course of the probability of correct recognition, as controlled by the successively added speech sounds, has any diagnostic value with respect to the type of degradation of the speech signal.

2. Method

A set of 40 Dutch polysyllabic words was selected, with frequencies of usage of 10 or more per 720.000 words in the Uit den Boogaart word frequency count (Uit den Boogaart, 1975). Optimal spoken realisations of these words by a speaker of standard Dutch were recorded and stored on disk in digital form (pcm, 12 bit per sample, 10 kHz sampling frequency). From each word token an initial fragment was isolated, corresponding to the beginning of the word, and containing several speech sounds. This fragment was chosen such that it was long enough to successfully apply LPC analysis and resynthesis, and short enough to ensure a low probability of correct recognition. Further versions of the same word token were produced by adding segments of speech corresponding to successive speech sounds to the initial fragment. This was done under visual and auditory control. An example of a phonetic transcription of consecutive fragmentary word tokens of one word, in this case the word AUTORITEIT (Engl. AUTHORITY) is:

1. [oto], 2. [otor], 3. [otori], 4. [otorit] 5. [otoritɛi], 6. [otoritɛit].

All 40 sets of word fragments were prepared in four speech qualities:
1. the original digital recording, using 120,000 bits per second;
2. vocoder speech, obtained with an LPC-to-formant analysis-resynthesis system, using 16,000 bits per second (Cf. 't Hart, Nooteboom, Vogten and Willems, (1982);
3. idem, with further data reduction by parameter quantisation to 4,000 bits per second;
4. idem, with still further reduction to 1,000 bits per second.

From these 40 sets of word fragments in the four speech qualities, four stimulus tapes were prepared. Each tape contained four groups of ten words, each group in a different speech quality. Each group of ten words appeared in a different speech quality on each of the tapes. The order of speech qualities on each tape varied randomly from one word to the next.

Each tape was played over headphones to a different group of five listeners, who were tested individually. After the presentation of each fragment listeners were encouraged to guess and say aloud the word from which the current fragment was taken. If not able to guess, they were asked to repeat aloud the fragment heard. After each correct guess the experimenter switched to the next set of word fragments. Stimuli and responses were recorded on two separate tracks of a magnetic tape for later analysis.

3. Results

The results presented here will be limited to probabilities of correct recognition as a function of the number and kind of added segments. Probability of correct recognition as a function of the number of speech segments added to the initial word fragment, for the four speech qualities separately and averaged over all words and all subjects, is given in Fig. 1.

The difference between each pair of curves is significant ($p < 0.05$) on a sign test applied to estimated means for individual words in different conditions. As expected, the number of audible segments necessary for correct recognition systematically increases with decreasing speech quality.

In search for diagnostic indications in our data, we have calculated the relative contribution of consonant and vowel segments to correct recognitions. The proportion of the total number of correct recognitions occurring immediately after adding a vowel segment, and the proportion occurring immediately after a consonant segment, in the four speech qualities, is plotted in Fig. 2. We see that with decreasing speech quality the relative contribution of vowel segments increases at the cost of consonant segments.

We also investigated the relative contribution of stressed and unstressed syllables to recognition. For this purpose we focused on those 27 of the 40 words in which the initial word fragment did not contain the lexically stressed syllable. For each of those words we numbered the added segments,

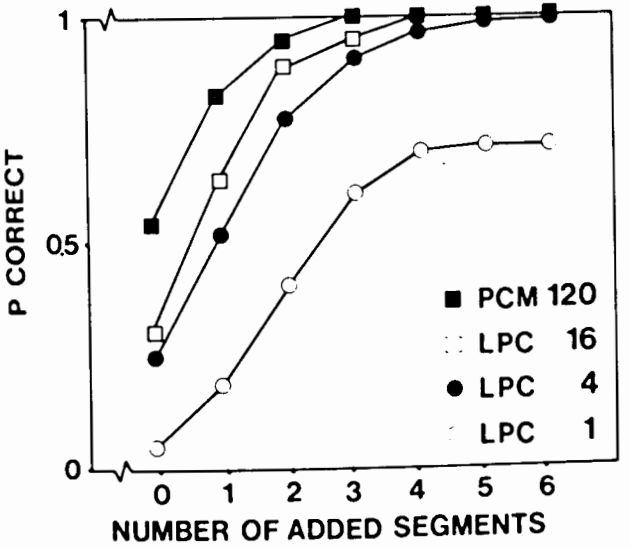

Figure 1. Probability ot correct word recognition as a function of the number of sound segments added to the initial word fragment, for four speech qualities.

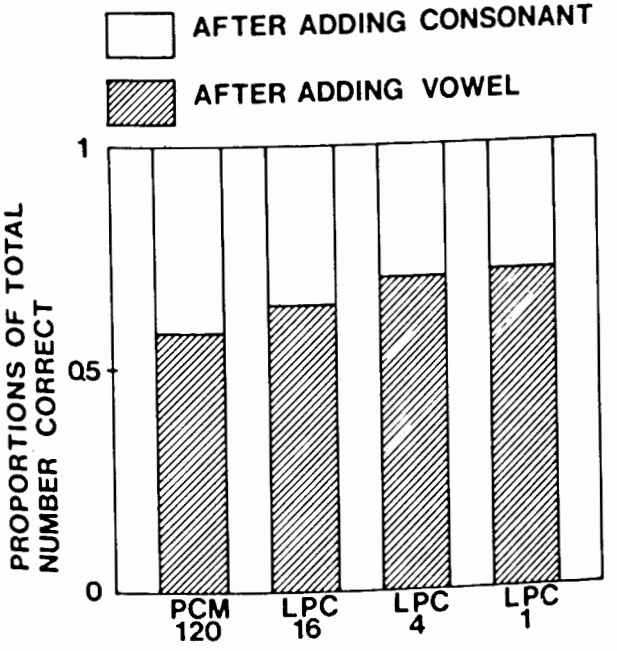

Figure 2. Proportions of total number correctly recognized words after adding a vowel or a consonant segment, for four speech qualities.

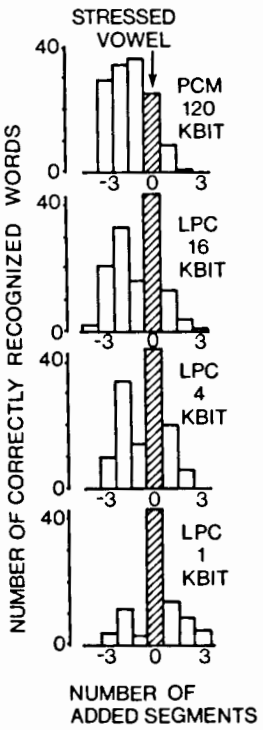

Figure 3. Frequencies of correct word recognition as a function of the position of the added segment. This position is taken relative to the position of the lexically stressed vowel.

starting with 0 for the vowel of the stressed syllable, negatively towards earlier and positively towards later segments. We then made frequency distributions of correct recognitions over the numbers of added segments. These are presented in Fig. 3. Obviously, as speech quality decreases correct recognition becomes more and more dependent on the availability of the vowel of the lexically stressed syllable.

4. Discussion and Conclusion

The results of this experiment show that the 'gating paradigm' can fruitfully be applied to the problem of measuring differences in speech quality. It proved possible to find significant differences between the four speech qualities used, with 40 words and only a few listeners per word, suggesting that measurement of relative speech qualities can be fairly easy and fast, given the availability of prepared sets of word fragments. The discriminative power of the test compares favourably with an adaptation of the Nakatani and Dukes (1973) test, as applied to approximately the same speech qualities by Vogten (1980). As exemplified in the results section, a simple analysis of the data distribution may give useful indications which parts of the speech signal are most seriously damaged in each speech quality.

References

Grosjean, F. (1980). Spoken word recognition and the gating paradigm. *Perception and Psychophysics,* **28**, 267-283.

Hart, J. 't, Nooteboom S.G., Vogten, L.L.M. and Willems, L.F. (1982). SPARX: manipulation of speech sound. *Philips Technical Review,* **40**, 134-145.

Marslen-Wilson, W.D. (1980). Speech understanding as a psychological process. In: J.D. Simon (Ed.) *Spoken Language Generation and Recognition.* Dordrecht: Reidel.

Nakatani, L.H. and Dukes, K.D. (1973). A sensitive test of speech communication quality. *Journal of the Acoustical Society of America,* **53**, 1083-1092.

Uit den Boogaart, P.C. (1975). *Woordfrequenties in geschreven en gesproken Nederlands.* Utrecht: Oosthoek, Scheltema en Holkema.

Vogten, L.L.M. (1980). Evaluation of LPC formant-coded speech with a speech interference test. *IPO Annual Progress Report,* **15**, 33-41.