As text-to-speech systems develop it becomes necessary to compare various solutions and to evaluate whether a change in the synthesis procedure has an effect on the listener's attitude to the system. Because there are no physical measurements that result in indices that quantify perceptual attributes of synthesized speech, psychophysical tests need to be used. The present study assesses the effectiveness of measuring listeners' impressions of synthesized speech using a magnitude estimation task. In particular, this study focuses on acceptability (i.e. the overall users' satisfaction with the communication) and understanding (i.e. the linguistic message), and naturalness (i.e. how much the system sounds like a normal human talker). The study consists of three experiments which are described further in the text. Depending on the experiment, four or seven synthesizers were evaluated in one or two types of external distortion (noise), as well as in quiet.

GENERAL METHODOLOGICAL CONSIDERATIONS

One hundred and fourteen normal-hearing college students participated as subjects. For obtaining magnitude estimations semantically-correct sentences (5 - 8 words in length) were extracted from a novel in a random manner. The same sentences, as well as semantically-anomalous sentences, were used for testing speech recognition. The latter had nearly the same syntactic structure as the semantically-correct sentences.

Seven different text-to-speech synthesis procedures ("systems") for the French language were used. Five of them (the "CNET" group) were defined by using the diphone synthesis algorithm developed at the Centre National d'Etude de Telecommunication. The parameters varied were the sampling frequency (12 kHz or 16 kHz), prosody (L = prosody used in reading, or P = prosody used by TV announcers), and voice (M = male, F = female). The five systems were as follows: C1: 12 kHz, L, M; C2: 12 kHz, P, F; C3: 16 kHz, P, M; C4: 16 kHz, L, M; C5: 16 kHz, L, F. The differences between the stimuli provided by these systems range from very small (systems differing in one synthesis parameter only) to reasonably large (systems differing in all three parameters). In order to evaluate whether rating of a system depends on the perceived distances from other systems used, the remaining two systems were intended to be substantially different from the CNET group and from each other. They were the system Infovox SA 201/PC and the system L'Info.

Depending on the experiment the synthesizers were evaluated in one or two types of distortion (external noise), as well as in quiet. The intent was to vary the comprehension difficulty from "easy" to "medium" to "difficult." The internal imperfections of the system will also be referred to as distortion. Therefore, the first type of distortion (D1) was the speech in quiet condition. The second type of distortion (D2) was created by band-pass noise (100 to 1000 Hz) at -4 dB S/N ratio. For the third type of distortion (D3) another band-pass noise (100 to 2000 Hz) was presented at -5 dB S/N ratio.

The magnitude estimation procedure consisted of presenting, in an attitude study, 124 sentences to the listener and asking him to assign a number that was equal to the magnitude of his impression of the judged attribute. A listening condition was defined as a combination of a synthesizer and distortion type. Depending on the group and the experiment, three or four judgments per condition were obtained from each subject. The first series of judgments was considered as preliminary, and the latter two (or three) were used for computing the appropriate group statistics. The preliminary set of judgments will be referred to as the "nominal" practice to differentiate it from the extensive practice that some subjects were exposed to.

The sentences were allocated randomly to each condition, i.e. to each synthesizer-distortion combination. Each subject had a different randomization of condition order. The sentences did not repeat, and, with the exception of group D in Experiment I where the sentence familiarity effect was investigated, the subject heard them for the first time. The results were calculated as "absolute," or "relative." The absolute value for a condition was calculated by finding the geometric mean across subjects and condition repetitions. To obtain the relative value the raw estimations of each individual were first divided by the geometric mean of all that individual's estimations. The arithmetic mean of these normalized estimations, across subjects and condition repetitions, represented the relative value of the condition. The correlation between the relative and absolute judgments was found to be high. It ranges from 0.93 to 0.997 (mean value 0.98).

EXPERIMENT I

In this study it is investigated whether the magnitude estimation judgments vary with the number and range of test conditions, or whether they depend on the subject's familiarity with the test material. It is also studied whether the magnitude estimation scales are practice invariant.

Fifty four subjects were divided into four
experimental groups. All groups gave magnitude estimations of acceptability. Group A was presented with the complete set of experimental conditions. That is, the subjects listened to all combinations of seven synthesizers and three distortion types (21 conditions in the total).

Groups B, C, and D differed from group A in that the subjects were presented with only eight experimental conditions. The eight conditions were defined as combinations of the two smaller distortion types (D1 and D2) and four systems of the CNMT group (C1, C2, C3, and C5). Therefore, both the number of conditions and the range of differences for these groups were substantially smaller than those for group A. In other respects group B did not differ from group A; therefore, the differences in their results, if any, could be attributed to the effects of the condition set size and range. Group C differed from group B in that the subjects in group C obtained extensive practice in all test conditions. Group D differed from group C in that subjects in group D were tested with the familiar test material; i.e., with the same sentences presented already eight times in the extensive practice period. Thus, the differences between groups B and C, and C and D should yield information on, respectively, the effect of practice and the familiarity with the test material.

An analysis of variance indicated that the effect of the experimental group was significant. The differences between groups A and B, and between groups B and C did not reach statistical significance. The difference between groups C and D was significant. Therefore, neither a condition size/range effect nor a practice effect was demonstrated. However, there seems to exist an effect of familiarity with the test material.

In this analysis a hypothesis was made that the absolute judgments are on an absolute scale, i.e., that different groups of subjects give the same mean numbers for the same listening conditions. This is, for example, the case with the loudness judgments. Thus far, however, in the case of magnitude estimates, this is only a hypothesis. Therefore, the results discussed above may only be artifacts due to the hypothesis being false. However, that for these artifacts to occur the effects of each, the condition set size/range variable, and the practice variable, should have been almost exactly counterbalanced by the differences in scales that each group used. This does not seem very likely, and it could be hypothesized that the judgments of acceptability are, indeed, on an absolute scale. These data, however, could not provide for an accurate test of the hypothesis.

To explore the questions of condition set size/range, practice, and familiarity effect, without confounding the analysis with the hypothesis of the absolute nature of the judgments, the relative results are analyzed. In Fig. 1 the relative magnitude estimations of acceptability for group A and group B are given. In order to better appreciate the psychological distances between the conditions (i.e., to represent equal stimulus ratios by equal intervals on the ordinate) the logarithms of the relative ratings are used. The relative ratings were adjusted so that for the two groups they agree for the best and worst condition that they had in common. The graph illustrates that not only was the condition set size larger for group A than for group B, but that also the psychological acceptability distances among the conditions of group B were substantially smaller than the distances among the conditions in group A. The relative results between groups A and B are highly correlated (0.96), as do the absolute results, that there was no effect of the condition set size and range.

In conclusion, it appears that effects of the condition set size, range, and practice, do not have an influence on the magnitude estimations of
acceptability. However, the familiarity with the test material is important for the judgments of acceptability. Therefore, if the same listeners are used repeatedly in testing synthesizers, it is necessary to create a mechanism for generating a quasi-infinite pool of test material. Finally, the data provide some evidence, albeit inconclusive, that the magnitude estimates of intelligibility are on an absolute scale.

FIG. 3. As Fig. 2 but for group C (squares) and group D (plusses).

EXPERIMENT II

In this experiment the relationship between the "objective" measures of speech intelligibility (proportion of words understood correctly) and the "subjective" measures (magnitude estimations) is examined. If these measures are highly related, then not only can the more efficient magnitude estimations be substituted for the less efficient objective tests, but this will also attest to the possible validity of the magnitude estimations. Further, the relationship between the speech recognition scores on semantically-correct and semantically-anomalous sentences is investigated.

Forty-five subjects were divided into three experimental groups (A, B, and C) of 15 subjects each. Group A gave magnitude estimations of intelligibility, while groups B and C were used to obtain speech recognition scores. Group B was tested with semantically-correct sentences, while group C was tested with semantically-anomalous sentences. All groups were presented with the complete set of experimental conditions (21). The speech recognition test structure paralleled as closely as possible that of the magnitude estimation test structure. The tests were scored as proportion of words repeated correctly. The proportion of correct responses for a condition was calculated by finding the arithmetic mean across subjects and condition repetitions.

The correlation between the group speech recognition scores on the semantically-correct and semantically-anomalous sentences across all conditions of distortion type and synthesizer type was high (0.98). In the three types of distortion, D1, D2, and D3, the correlations were also high; respectively, 0.99, 0.98, and 0.98. That a high correlation was seen even in select (D1), was partly due to the fact that the performance of the synthesizers as a group, was less an ideal (mean score on the semantically-correct sentences was 88.4%). Thus, the effect of the 100% ceiling was minimal.

The high linear correlation between these two objective measures of speech intelligibility suggests that the speech synthesis does not influence listeners' semantic processing abilities independent of the influences already present at the segmental and prosodic levels. Thus, semantically anomalous, but syntactically-correct, sentences could be used for assessing speech intelligibility. In relatively easy communication situations they are less subject to the ceiling effect than semantically-correct sentences.

A correlation analysis was performed on the relative magnitude estimations of intelligibility and the objective speech recognition scores. The correlation coefficients for each distortion type separately, and across all distortion types are high. They range between 0.89 and 0.94. This implies that the magnitude estimations are valid measures of speech intelligibility. This suggests that the magnitude estimates of other speech quality attributes may also be valid. Further, these results justify the use of more efficient magnitude estimation measures of speech intelligibility instead of the objective speech recognition scores.

EXPERIMENT III

In this experiment it is studied how the magnitude estimation scales of acceptability, naturalness, and intelligibility vary with the severity of external distortion. Further, it is investigated whether there are important dependencies among acceptability, naturalness, and intelligibility. In particular, it is studied whether acceptability could be sufficiently explained by naturalness and intelligibility.

Forty-five subjects were divided into three experimental groups of 15 subjects each. Each group had a different task. That is, the subjects from one group judged only acceptability, the subjects from another group judged only intelligibility, and the subjects from the third group judged only naturalness. In this way carry-over from one type of judgement to another was prevented. All groups were presented with the complete set of 21 experimental conditions, that is with all combinations of synthesis types (7), and distortion types (3).

Fig. 4 shows the absolute magnitude estimations of acceptability of the seven synthesis systems in the distortion types D1, D2, and D3. An analysis of variance of these results indicates that both the main effect of the synthesizer type and distortion type are significant. The analysis also indicates a significant interaction between the ratings of the synthesizers and the distortion type. This is further corroborated by a correlation analysis on the relative ratings of the synthesizers. The correlation coefficients between the ratings in the distortion types D1 and D2, D1 and D3, and D2 and D3 are, respectively, 0.91, 0.46, and 0.67. Thus, the relative acceptability of the system changes with communication condition.

Fig. 5 shows the absolute magnitude estimations of intelligibility of the seven synthesis systems in the distortion types D1, D2, and D3. An analysis of variance of these results indicates that both the main effect of the synthesizer type and distortion type are significant. The interaction between the ratings of the synthesizers and the distortion type does not reach the 0.05 level of statistical significance. The correlation coefficients between the relative ratings in the distortion types D1 and D2, D1 and D3, and D2 and D3 are relatively high, and, respectively, 0.93, 0.82, and 0.93. This finding, however, may only be characteristic of the
distortion types and synthesizer systems used in this study. Thus, it cannot be generalized. In the case of the significant interaction, as was the case with acceptability, the finding is generalizable. That is, the fact that there exist some synthesizers and some distortion types that interact requires that when any other synthesizer, and any other distortion type, with unknown interaction are assessed, the evaluation of the synthesizer must be performed in the actual condition of distortion.

Fig. 6 depicts the absolute magnitude estimations of naturalness of the seven synthesis systems in the distortion types D1, D2, and D3. An analysis of variance of these results indicates that both the main effect of the synthesizer type and distortion type are significant. The correlation coefficients between the relative ratings in the distortion types D1 and D2, D1 and D3, and D2 and D3 are, respectively, 0.81, 0.74, and 0.83. Thus, the relative naturalness of the system changes with communication condition. While it is not unexpected that the main effect of the distortion type was significant, it is surprising that the ratings in the distortion type D2 were significantly higher than the ratings in quiet (distortion type D1). It appears that the presence of some background noise (but not too much of it as was the case in the distortion type D2) made the synthesizers sound more natural.

In general, the data indicates that the magnitude estimates of synthesizer performance should be done in the actual conditions of use. The relative ratings of acceptability, intelligibility, and naturalness may change with the distortion type. The absolute ratings of acceptability and intelligibility decrease with the increase of the external disturbance, while the absolute ratings of naturalness may, under some conditions, increase.

A correlation analysis indicates that acceptability, intelligibility, and naturalness are highly intercorrelated in each distortion type. Similarly, the accuracy of predicting acceptability based on intelligibility only, does not demonstrate a significant improvement if naturalness is included in the linear prediction equation. These results are somewhat surprising. Normally, it could be expected that in difficult listening conditions acceptability would be highly correlated with intelligibility, and less so with naturalness. In the easy listening conditions an opposite relationship may be anticipated. The failure of this experiment to demonstrate these relationships may be due to one or more of the following three factors. First, the synthesizers that were tested may have differed only in the realization of acoustical elements correlated with both naturalness and intelligibility. Second, the intelligibility of the synthesizers that were tested was far from ideal (mean score on the semantically-correct sentences was 88.4%). Thus, it is possible that even the D1 conditions belonged to the same class of conditions ('difficult') as did the D2 and D3 conditions. Third, the possibility that the observed results are characteristic of the synthesis systems and distortion types in general cannot altogether be ruled out.

Finally, it should be noted that acceptability was well explained by naturalness and intelligibility (r = 0.93). Thus, it appears that these two attributes of synthesizer performance are sufficient to describe its overall acceptability to the user. What remains under some doubt is whether indeed both of them are necessary.

FIG. 4. Absolute magnitude estimations of acceptability of the synthesis systems in the distortion types D1 (squares), D2 (pluses), and D3 (diamonds). The codes C1 to C5 are used for the systems of the CNET group, the code IV for the system Infovox, and the code LI for the system 1'ICophone.

FIG. 5. As Fig. 4. but for the absolute magnitude estimations of intelligibility.

FIG. 6. As Fig. 4. but for the absolute magnitude estimations of naturalness.

ACKNOWLEDGMENTS
This research was made possible by a grant from the EEC Esprit project.