Is synthetic speech just degraded speech or is it processed as a specific perceptual space? The identification responses to 8 phonetically balanced lists of ten sentences each, using several syntactic structures, were studied for four sets of stimuli (natural speech, LPC speech, synthesis by diphones using two text-to-speech systems). All the stimuli were intensity equalized, then degraded by a masking pink noise. Phonetic and prosodic cues effects were strong, while the effect of syntax was weak. The choice of sentence identification strategy depends on the natural vs synthetic nature of the speech used and on SNR: a step-by-step decoding for impoverished synthetic speech and a SNR below 8 dB, backward lexical interpretation for natural speech or a low noise. Acoustic cues redundancy and masking noise level impose the choice of specific cognitive processing modalities.

In the case of spoken language, sentence perception and comprehension imply the interaction of both acoustic and linguistic sources of knowledge to identify word boundaries, select word candidates and construct a meaningful sequence. According to identification tasks using a gating paradigm in which signal duration is varied /3, 4, 9/, data support the assumption of a parallel and interactive processing of acoustic-phonetic information and of syntactic-semantic information provided by the sentence context. It is the redundancy of lower-order and higher-order sources of information which can explain the listener's ability to understand speech even under degraded conditions. But the redundancy of acoustic-phonetic cues by themselves is also of importance. It is possible to evaluate its weight by comparing sentence recognition performance for natural speech and synthetic speech of different qualities.

Previous research has demonstrated that synthetic speech is more difficult to recognize than natural speech /8/. This is perhaps due to what Nusbaum and Pisoni /7/ call the "noisy speech" hypothesis i.e. the fact that acoustic structure of synthetic speech is somehow degraded, as in the case of natural speech in noise. But according to the "impo verished speech" hypothesis, the rather bad performance for synthetic speech corresponds to a specific cognitive processing. Listeners must adapt their perceptual and identification strategies to a signal which is in its nature different from natural speech: they have to build a new perceptual space.

The present experiment aims at studying how naive listeners, without a previous knowledge of synthetic speech, can manage to understand sentences with different degrees of syntactic complexity, either natural or digitized, or generated by good vs. low-cost text-to-speech systems. Moreover stimuli were degraded by adding varying amounts of pink noise. The main hypothesis is that the level of performance and kind of errors will be linked to the quality of the sets of stimuli i.e. to the characteristics of the potential perceptual space. In any cases they will be significantly different for natural and synthetic speech. Another assumption bears on the effect of syntactic and semantic complexity. As speech becomes less intelligible, according to either to its quality or to speech-to-noise ratio (SNR), listeners will rely more heavily on linguistic structure, so that easy-to-parse sentences would be better understood than less predictable ones, specifically as the quality of synthetic speech becomes worse /7/. Finally, following the researches on synthetic speech training /2/, it can be hypothesized that the results will improve from the first to the second session.

Speech materials and systems
Eight phonetically balanced lists of ten sentences each, covering a range of syntactic structures and semantic degrees of plausibility /1/, were read by a trained female speaker, with a neutral intonation and a 4.27 syllables/second speech rate. The first set
of stimuli, A1, consisted of these naturally spoken sentences. Audio tapes of the original sentences were replayed at 16 kHz (16 coefficients), digitized by a linear prediction coder and stored on disk by a PDP-11/34 computer. This second set of materials will be referred to as A2. The two other sets were generated using synthesis by diphones according to two text-to-speech systems. The high-quality one, A1, was processed with all features as described in Section II on a PDP-11/34 computer, and generated from a dictionary recorded by a male speaker at a 3.42 syl/sec. speech rate. Prosody was a good approximation of natural speech. The last set of stimuli, A4, was processed by a low-cost computer. This second set of materials will be referred to as A. The two other sets were generated using a diphone dictionary recorded by a female speaker at a 3.18 syl/sec. speech rate. This dictionary was implemented on a micro-processor (2K6 sec period). Some rough prosodic markers were added.

Procedure and subjects

Half of the subjects who saw the stimulus equally at 71/72 dB. The stimuli were masked by pink noise the intensity of which decreased from trial to trial. In the first trial, SNR were of 2 dB for natural speech, +4 dB for LPC speech and high-quality system, +8 dB for low-cost system. These values were chosen so that no correct response could be given at the first presentation. At each of the 6 successive presentations, the level of noise was diminished by 2 dB steps for the first 3 presentations, then 3 dB steps for the next synthetically speech. Four groups of 5 subjects each participated in the experiment according to 2 sessions, at an interval of 5 days. All groups were given the same recognition task. The amount of noise they had understood after each presentation of each session. One half of the stimuli were counterbalanced, and the systems were crossed with the lists according to a Latin square design. For each group the factorial design was as follows:

\[ S_1 \times A_1 \times D_1 \times S_2 \times A_2 \times D_2 \times S_3 \times A_3 \times D_3 \]

(S: subjects, L lists, A: systems, D: test session, S: SNR). Speech-to-noise ratio at the identification threshold was measured for correct responses. The correct response percentages for each list, sentence, and system were determined, and SNR were evaluated for correct responses as a function of systems and sessions. A preliminary analysis of variance was performed on the data, with a significant result for the main effect of SNR on correctness. As a result, the data were grouped according to the factor SNR. The data were then analyzed separately for each SNR level. The analysis of variance was performed on the data, with a significant result for the main effect of SNR on correctness. As a result, the data were grouped according to the factor SNR. The data were then analyzed separately for each SNR level. The analysis of variance was performed on the data, with a significant result for the main effect of SNR on correctness. As a result, the data were grouped according to the factor SNR. The data were then analyzed separately for each SNR level. The analysis of variance was performed on the data, with a significant result for the main effect of SNR on correctness. As a result, the data were grouped according to the factor SNR. The data were then analyzed separately for each SNR level.
graded by noise. On the contrary, our results agree with the definition of synthetic speech as "impoverished speech" /7/, different in its nature from natural speech. They support the conclusion that the differences of intelligibility between natural and synthetic speech are related to the characteristics of speech signal. Different generating systems offer different patterns of cues to listeners. So listeners must construct and process several "perceptual spaces". The three synthetic speech systems generally present the same kind of confusion errors, more or less frequent depending on the quality of the system. Furthermore, two kinds of processing strategies can be hypothesized: a step-by-step decoding strategy and a global comprehension strategy. But further research is needed to better understand how perceptual spaces are built, what their consistency is, and how their processing can be improved.

CONCLUSIONS

Impoverishment of speech by a pink noise varied mainly as a function of the system from which signal was generated. Relative weakness and lack of stability of sentence effect suggest that perceptual processing, in this experiment, has borne mainly on acoustic-phonetic cues, and secondly on prosodic segmentation cues. Listeners relied more on acoustic than on specifically linguistic information. Higher-order information was used, as demonstrated by the occurrence of backward lexical identification mechanisms; but its effect depends on the main effect of the quality of the system. Our results agree with the conclusion of Nusbaum and Pisoni /7/, "the differences in perception of natural and synthetic speech are largely the result of differences in the acoustic-phonetic structure of the signals" (p. 239). However, unlike them, we found that linguistic context becomes more important as the quality either of speech or of listening gets better, as is the case when one examines error restoration as well as identification thresholds. Furthermore, acoustic information is all the more processed as either speech quality or SNR are worse. In such bad listening conditions, subjects process the signal in a step-by-step fashion, more clearly so for synthetic speech than for natural speech. Dissymmetry between responses is sufficient to rule out the hypothesis that synthetic speech is equivalent to natural speech degraded by noise. On the contrary, our results agree with the definition of synthetic speech as "impoverished speech" /7/, different in its nature from natural speech. They support the conclusion that the differences of intelligibility between natural and synthetic speech are related to the characteristics of speech signal. Different generating systems offer different patterns of cues to listeners. So listeners must construct and process several "perceptual spaces". The three synthetic speech systems generally present the same kind of confusion errors, more or less frequent depending on the quality of the system. Furthermore, two kinds of processing strategies can be hypothesized: a step-by-step decoding strategy and a global comprehension strategy. But further research is needed to better understand how perceptual spaces are built, what their consistency is, and how their processing can be improved.

REFERENCES