V. CONCLUSIONS

In the first tests that were conducted with the speech synthesizer, the synthesized speech was judged to be unnatural in character although it was surprisingly intelligible. In initial tests that were conducted to determine the intelligibility of synthesized test words, approximately 75 percent of the synthesized words were correctly understood by the average listener. Since these words were in the form of unconnected speech, one would expect a relatively higher intelligibility for ordinary conversational speech. Furthermore the percentage articulation would be higher if the listeners had been given the opportunity of listening to the synthesized speech for a period of time and of becoming familiar with it.

The device for synthesizing speech which has been described here is useful primarily as a research tool for further study of types of standardized speech. Although speech has been synthesized by placing in tandem, sounds which were taken from the natural speech of a single individual, the device could be employed equally well for synthesizing speech that consists of arbitrarily selected sounds as its basic units. In either case, the principal advantage of this synthesizer is the element of control which it introduces in such a study in that the same module may be used any number of times in combination with any of the other modules under test. Further experimentation will undoubtedly indicate that speech modules other than those employed in the initial study will result in higher intelligibility.

THE JOURNAL OF THE ACOUSTICAL SOCIETY OF AMERICA VOLUME 25, NUMBER 5 SEPTEMBER, 1953

Some Experiments on the Recognition of Speech, with One and with Two Ears*

E. COLIN CHERRY
Imperial College, University of London, England, and Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, Massachusetts

(Received May 5, 1953)

This paper describes a number of objective experiments on recognition, concerning particularly the relation between the messages received by the two ears. Rather than use steady tones or clicks (frequency or time-point signals) continuous speech is used, and the results interpreted in the main statistically.

Two types of test are reported: (a) the behavior of a listener when presented with two speech signals simultaneously (statistical filtering problem) and (b) behavior when different speech signals are presented to his two ears.

1. INTRODUCTION

THE experiments described herein are intended as a small contribution to the solution of the general problem of the recognition of speech. They are designed to be essentially objective and behavioristic; that is, the "subject" under test (the listener) is regarded as a transducer whose responses are observed when various stimuli are applied, whereas his subjective impressions are taken to be of minor importance.

A great deal of work has been done relating to aural discrimination, mostly using two kinds of stimulus: (a) pure tones, which may be regarded as separable in frequency, and (b) acoustic "clicks," or impulses, considered as separable in time.[1] It is suggested that a third kind of discrimination is possible and amenable to experimental treatment, namely statistical separation. Speech signals form stimuli in this class, and we appear to possess powers of such discrimination. For example, we decide that a person is speaking English and not, say, French; again we can listen to one speaker when another is speaking simultaneously. These are acts of recognition and discrimination.

The tests to be described are in two groups. In the first, two different spoken messages are presented to the subject simultaneously, using both ears. In the second, one spoken message is fed to his right ear and a different message to his left ear. The results, the subject's spoken reconstructions, are markedly different in the two cases; so also are the significances of these results. Before examining such possible significance, it will be better to describe some of the experiments.

2. THE SEPARATION OF TWO SIMULTANEOUSLY SPOKEN MESSAGES

The first set of experiments relates to this general problem of speech recognition: how do we recognize

* This work, supported in part by the Signal Corps, the Air Materiel Command, and the U. S. Office of Naval Research, was carried out by the author at M.I.T. while there as a Visiting Professor under a Fulbright grant, and is presented with the kind permission of Professor J. B. Wiesner.

[1] M. R. Rosenzweig, Am. J. Physiol. **167**, No. 1 (October, 1951).

what one person is saying when others are speaking at the same time (the "cocktail party problem")? On what logical basis could one design a machine ("filter") for carrying out such an operation? A few of the factors which give mental facility might be the following:

(a) The voices come from different directions.
(b) Lip-reading, gestures, and the like.
(c) Different speaking voices, mean pitches, mean speeds, male and female, and so forth.
(d) Accents differing.
(e) Transition-probabilities (subject matter, voice dynamics, syntax. . .).

All of these factors, except the last (e), may, however, be eliminated by the device of recording two messages on the same magnetic-tape, spoken by the same speaker. The result is a babel, but nevertheless the messages may be separated.

The logical principles involved in the recognition of speech seem to require that the brain have a vast "store" of probabilities, or at least of probability-rankings. Such a store enables prediction to be made, noise or disturbances to be combatted, and maximum-likelihood estimates to be made. Shannon[2] has already reported that prediction is readily possible in the case of printed language, and has described experiments in which a subject is required to guess the successive letters or words of a hidden written message; our present experiments are somewhat analogous, but are carried out with speech, at normal rates of speaking.

Those holding the strict behaviorist view may rightly object that it is inadmissable to speak of "storage of probability-rankings in the brain," because these are not directly observable; the only probabilities which can be discussed are those of the subject's responses. Acknowledging this, we may turn the problem around from one of psychology to one of engineering and ask: On what logical principles could one design a machine whose reaction, in response to speech stimuli, would be analogous to that of a human being? How could it separate one of two simultaneous spoken messages? The tests described here merely purport to show that we ourselves have such power, with the suggestion that we can assess probability-rankings of words, phonemic sounds, syntactical endings, and other factors of speech.

In the first experiment the subject is presented with the two mixed speeches recorded on tape and is asked to repeat one of them word by word or phrase by phrase. He may play the tape as many times as he wishes and in any way. His task is merely to separate one of the messages. He repeats the various identified portions verbally, but is not allowed to write them down.

The following is one example of two messages, showing his reconstructions; the subject matters are markedly distinct in this case.

[2] C. E. Shannon, Bell System Tech. J. XXX, 50 (1951).

Message 1 (a) "It may mean that ~~our~~ religious convictions, legal systems and politics have been so successful in accomplishing their ~~ends~~ AIMS during the past two thousand years, that there has been no need to change our outlooks about them. Or it may mean that the outlook has not changed for other reasons. I ~~will~~ leave the first hypothesis BELIEVE IN for those who are willing to defend it, and choose the second. As the AND IN reader may have guessed, I am interested in learning how obsolete structure of languages preserves obsolete metaphysics."

Message 1 (b) "This very brief discussion will serve to give a slight indication of the really complex nature of the causes and uses of birds' colors, and may serve to suggest a few of the many possibilities that may underlie them. There is a very great opportunity here for close and careful observation of the habits of birds in a free state, with a view to shedding light on these problems But the observer, in interpreting what he sees, must ever be on his guard lest he lose sight of alternative explanation."

The phrases recognized have been underlined and error indicated by the sub-scripts. No transpositions of phrases between the messages occurred in this example; in other examples extremely few transpositions arose, but where they did they could be highly probable from the text. The next example illustrates this point (indicated by asterisks).

Message 2 (a) "He came ~~out of~~ nowhere special; a cabin like any other out FROM West. ~~His~~ folks were nobody special; pleasant, hardworking people like HE SPOKE TO many others. Abe was a smart boy but not too smart He could do a good day's work on the farm, though he'd just as soon stand around and talk. ~~He told~~ funny* stories; he was strong and kind. He'd never try to hurt you, or PROFESSIONAL TRAINING cheat you, or fool you. Young Abe worked at odd jobs and read ~~law~~ books at WAR night. Eventually he found his way into local politics. And it was then that people, listening* to his speeches, began to ~~know~~ there was something special LEADING POSITION IN THE WORLD NOTICE about Abe Lincoln. Abe talked about running a country as though it were THE something you could do. It was just a matter of people getting along. HE He had nothing against anybody, rich or poor, who went his own way and let GO the other fellow go his. No matter how mixed up things got, Abe made you feel that the answer was somewhere among those old rules that everybody knows: AMONGST no hurting, no cheating, no fooling."

Notice here the recognition in phrases, the highly likely errors and transpositions and the consistency of any initial grammatical mistake. Similar factors were observed in all the samples taken.

At the subjective level the subject reported very great difficulty in accomplishing his task. He would shut his eyes to assist concentration. Some phrases were repeatedly played over by him, perhaps 10 to 20 times,

but his guess was right in the end. In no cases were any long phrases (more than 2 or 3 words) identified wrongly.

Message 2 (b) "In attaining its present* position, the Institution
SPECIAL
has constantly kept before it three objectives - the education of men,
the advancement of knowledge and service to industry* and the nation
OTHERS
It aims to give its students such a combination of humanistic, scientific
and professional training as will fit them to take leading positions in
a world in which science, engineering and architecture are of basic importance.
THE
This training is especially planned to prepare students, according to their
HAS BEEN
desires and aptitudes, to become practicing engineers or architects, investigators,
business executives or teachers. The useful knowledge and mental discipline
AND
gained in this training are, however, so broad and fundamental as to constitute
an excellent general preparation for other careers.* Realizing that the
PEOPLE GETTING ALONG
Institution trains for life and for citizenship as well as for a career, its
TRAINING ASSOCIATIONSHIP (?)
Staff seeks to cultivate in each student a strong character, high ideals, and
a sense of social responsibility, as well as a keen intellect "

In a variation of the experiment the subject was given a pencil and paper, and permitted to write down the words and phrases as he identified them. Subjectively speaking, his task then became "very much easier." Times were shortened. It appears that the long-term storage provided assists prediction.

Numerous tests have been made, using pairs of messages of varying similarity. Some test samples consisted of adjacent paragraphs out of the same book. The results were consistently similar; the messages were almost entirely separated.

However, it was considered possible to construct messages which could not be separated with such a low frequency of errors. Such a test is described in the next Section.

3. INSEPARABLE SPOKEN MESSAGES. USE OF CLICHÉS OR "HIGHLY-PROBABLE PHRASES"

As a final test in this series, using the same speaker recorded as speaking two different messages simultaneously, a pair of messages was composed which could not be separated by the listening subject. The messages were composed by selecting, from reported speeches in a newspaper, 150 clichés and stringing them together with simple conjunctions, pronouns, etc., as continuous speeches. For example, a few of the clichés were:

(1) I am happy to be here today,
(2) The man in the street,
(3) Stop beating about the bush,
(4) We are on the brink of ruin,

and the like. The corresponding sample of one speech was as follows:

"I am happy to be here today to talk to the man in the street. Gentlemen, the time has come to stop beating about the bush—we are on the brink of ruin, and the welfare of the workers and of the great majority of the people is imperiled," and so forth.

It is remarkably easy to write such passages by the page.† Now a cliché is, almost by definition, a highly probable chain of words, and on the other hand the transition probability of one cliché following another specific one is far lower. The subject, as he listened to the mixed speeches in an endeavor to separate one of them was observed to read out complete clichés at a time; it appeared that recognition of one or two words would insure his predicting a whole cliché. But he picked them out in roughly equal numbers from both speeches; in such artificially constructed cases, message separation appeared impossible. The speeches were of course read with normal continuity, and with natural articulatory and emotional properties, during their recording.

It is suggested that techniques such as those described in the preceding sections may be extended so that they will shed light on the relative importance of the different types of transition probabilities in recognition. For instance, speeches of correct "syntactical structure" but with no meaning and using few dictionary words may readily be constructed. [Lewis Carroll's "Jabberwocky" is such an instance; similarly, "meaningful" speeches with almost zero (or at least unfamiliar) syntactical or inflexional structure (Pidgin English).] Again continuous speaking of dictionary words, which are relatively disconnected, into "meaningless phrases" is possible; the word-transition probabilities may be assessed *a priori*, with the assistance of suitable probability tables. Further experiments are proceeding.

4. UNMIXED SPEECHES; ONE IN THE LEFT EAR AND ONE IN THE RIGHT

The objective, and subjective, results of a second series of tests were completely different. In these tests one continuous spoken message was fed into a headphone on the subject's left ear and a different message applied to the right ear. The messages were recorded, using the same speaker.[1]

The subject experiences no difficulty in listening to either speech at will and "rejecting" the unwanted one. Note that aural directivity does not arise here; the earphones are fixed to the head in the normal way. To use a loose expression, the "processes of recognition may apparently be switched to either ear at will." This result has surprised a number of listeners; although of course it is well known to anyone who has made hearing tests. It may be noteworthy that when one tries to follow the conversation of a speaker in a crowded noisy room, the instinctive action is to turn one ear toward him, although this may increase the difference between the "messages" reaching the two ears.

† Comment upon this fact has appeared in the *New Yorker* under the name of Mr. Arbuthnot.

The subject is instructed to repeat one of the messages concurrently while he is listening[3] and to make no errors. Surprising as it may seem this proves easy; his words are slightly delayed behind those on the record to which he is listening. One marked characteristic of his speaking voice is its monotony. Very little emotional content or stressing of the words occurs at all. Subjectively, the subject is unaware of this fact. Also he may have very little idea of what the message that he has repeated is all about, especially if the subject matter is difficult. But he has recognized every word, as his repeating proves.

But the point of real interest is that if the subject is subsequently asked to repeat anything of what he heard in his other (rejected-message) ear, he can say little about it at all, except possibly that sounds were occurring.

Experiments were made in an attempt to find out just what attributes, if any, of the "rejected" message are recognized.

5. LANGUAGE OF "REJECTED" EAR UNRECOGNIZED

In a further set of tests the two messages, one for the right ear and one for the left, started in English. After the subject was comfortably repeating his right-ear message, the left-ear message was changed to German, spoken by an Englishman. The subject subsequently reported, when asked to state the language of the "rejected" left-ear message, that he "did not know at all, but assumed it was English." The test was repeated with different, unprepared listeners; the results were similar. It is considered unfair to try this particular test more than once with the same listener.

It was considered that a further series of tests might well indicate the level of recognition which is attained in the "rejected" ear, raising the questions, Is the listener aware even that it is human speech? male or female? and the like.

6. WHAT FACTORS OF THE "REJECTED" MESSAGE ARE RECOGNIZED?

In this series of tests the listening subjects were presented at their right-hand ears with spoken passages from newspapers, chosen carefully to avoid proper names or difficult words, and again instructed to repeat these passages concurrently without omission or error. Into their left ears were fed signals of different kinds, for different tests, but each of which started and ended with a short passage of normal English speech in order to avoid any troubles that might be involved in the listener's "getting going" on the test. The center, major, portions of these rejected left-ear signals thus reached the listener while he was steadily repeating his right-ear message.

Again no one listening subject was used for more than one test; none of them was primed as to the results to be expected. The center, major, portions of the left-ear signals for the series of tests were, respectively:

(a) Normal male spoken English—as for earlier tests.
(b) Female spoken English—high-pitched voice.
(c) Reversed male speech (i.e., same spectrum but no words or semantic content).
(d) A steady 400-cps oscillator.

After any one of these tests, the subject was asked the following questions:

(1) Did the left-ear signal consist of human speech or not?
(2) If yes is given in answer to (1), can you say what it was about, or even quote any words?
(3) Was it a male or female speaker?
(4) What language was it in?

The responses varied only slightly. In no case in which normal human speech was used did the listening subjects fail to identify it as speech; in every such instance they were unable to identify any word or phrase heard in the rejected ear and, furthermore, unable to make definite identification of the language as being English. On the other hand the change of voice—male to female—was nearly always identified, while the 400-cps pure tone was always observed. The reversed speech was identified as having "something queer about it" by a few listeners, but was thought to be normal speech by others.

The broad conclusions are that the "rejected" signal has certain statistical properties recognized, but that detailed aspects, such as the language, individual words, or semantic content are unnoticed.

7. SIMILAR MESSAGES IN THE TWO EARS, BUT WITH TIME DELAY BETWEEN THEM

Subjectively speaking, the effect of listening normally, with both ears, to a single message is a very different sensation from that of listening with one ear to one of two different messages as in the earlier tests. This raises the question of how we correlate the signals reaching our two ears so that we are able to decide to listen either to both at the same time (when identical or "correlated") or only to one, rejecting the other.

This question suggested the following experiment. Suppose we apply identical messages to the two ears of a listening subject, but with a very long delay between them. What will be the effect if this delay is steadily reduced, as the message proceeds, until eventually the two ears are stimulated simultaneously and identically?

Preliminary experiments suggest that the basis of correlation (using the word in the popular not the mathematical sense) of the messages reaching the two ears depends upon the magnitude of the delay between the ears. When this is very short, of the order of milliseconds, there will exist a considerable connection between the actual sounds, or their spectra; but with longer delays, of the order of seconds, the relation is

[3] D. E. Broadbent, J. Exptl. Psychol. 43 (April, 1952).

more a semantic one, or one of word and phrase identification.

The following experiment was carried out with a number of subjects. A long passage of speech was recorded on magnetic tape and subsequently run through two reproducing machines in cascade, with a length of tape between them. The subject, who was unprimed as to the nature and purpose of the experiment, was instructed in exactly the same way as in the earlier experiments; namely, he was asked to repeat the message reaching his right ear, without omission or error. As he was doing this the two machines were slowly pushed together, reducing the delay between the ears. At some stage the subject would exclaim: "My other ear is getting the same thing" or some equivalent remark. Some of them said nothing until asked afterwards and then stated the word or words first recognized as being the same. Nearly all subjects reported that they had recognized words or phrases, at some stage, in the rejected ear message as being the same as those in the accepted ear message.

The surprising thing here is that such words were recognized at all, because in earlier tests, using different texts for the two ears, not a single word of the rejected ear was identified. The delay at which recognition first occurred in the present tests varied considerably between the different listeners acting as subjects but mostly lay between 6 sec and 2 sec.

Experiments of a similar nature, but using very short delays of the orders of milliseconds or tens of milliseconds, are not reported here in connection with the present study. They are of interest mainly for the subjective effects produced.

8. THE SWITCHING OF ONE MESSAGE PERIODICALLY BETWEEN THE TWO EARS

This experiment was suggested by the results of earlier ones described in Secs. 4, 5, and 6. When listening to and repeating concurrently a message received in one ear while a different message is being presented to the other ear, it is found that a very short time interval is required to transfer the attention from the one ear to the other. Thus it was thought that, if a single message was switched between the ears at approximately the time period of this reaction time (not under the control of the listening-speaking subject), his recognition facility might be completely confounded and he would be unable to repeat the words.

A long sample of English speech was recorded on tape and subsequently applied to the right or left headphone of the subject, alternately, by an automatic switch which could be thrown (a) randomly and (b) periodically, at any required rate. When the switching speed was very slow (say a 1-sec period) the subject repeated 100 percent correctly; when very fast (say 1/20–1/50 sec period) most subjects repeated the majority of the words, though they varied in their ability considerably, reporting that they listened as though to both ears simultaneously. The point that matters is that an optimum period of switching could be found at which the fraction of words repeated by the subjects was a minimum. The flatness of this minimum varied between the subjects; the approximate average value of the minimum switching rate was $\frac{1}{6}$–$\frac{1}{7}$ sec, for a complete cycle of switching.

Somewhat surprisingly, little difference in the results was found between the uses of random and periodic switching; so the former was abandoned. The variations between the subjects in their abilities, the flatness of the minima, and other factors tended to make such experiments rather inconclusive. Instead, therefore, a method of switching was sought which could virtually stop any subject repeating any of the words. It was found that if, while the reversing switch was in operation, a very short gap of silence was introduced, the effect upon the subject's responses was most marked. The switching cycle was thus: right ear/silence/left ear/silence—periodically, at about 6 to 7 cps. The silence interval needed to be no greater than 10 msec.

A comparison measurement was made with each subject. Firstly, the ear-phone signals were not reversed, though the silence gap was introduced, the subjects thus listened to both ears, with the periodic (<10 msec) interval as interruption. Word scores were 95 to 100 percent correct. Then the reversal of the earphones was introduced; the word scores fell to less than 20 percent correct.

It may be considered that these results might be accounted for by the inherent noise introduced by the switching interruption of the speech; there are several factors which assist in denying this.

(a) The noise is at extremely low level when the switching rate is as slow as 6 to 7 per sec.

(b) A subject might get a high score with a silence gap of <1 msec but this would inevitably fall if the gap was opened. The noise is substantially unchanged.

(c) Miller and Licklider's results of experiments[4] carried out with periodically interrupted speech (both ears simultaneously) show that a 6-cps interruption of 50 percent of the time, that is, square-wave modulation of the speech, gave a word-articulation score as high as 75 percent; the noise introduced presumably being much the same as in our present experiment. The test material was somewhat different in their case, being individual monosyllabic words, not connected speech.

ACKNOWLEDGMENTS

Acknowledgment of the very great assistance offered by many patient subjects is gratefully made. The author wishes also to thank Professor J. B. Wiesner, Massachusetts Institute of Technology, and Professor Willis Jackson, Imperial College, London, for their assistance in affording the necessary facilities.

[4] G. A. Miller and J. C. R. Licklider, J. Acoust. Soc. Am. 22, 167 (1950).