Proc. 5th int. Congr. phon. Sci., Münster 1964, pp. 252–258
(S. Karger, Basel/New York 1965).

Bell Telephone Laboratories

On the Motor Theory of Speech Perception

By Peter B. Denes, Murray Hill, N.J.

Our understanding of how we perceive speech has, in many ways, decreased rather than increased over the years. The better understanding of vocal tract acoustics and advances in instrumentation in recent times, seem to have uncovered as many new problems about the process of speech recognition as they have solved. Years ago, speech recognition was thought to be a simple process in which distinct and unique acoustic-auditory features are interpreted as specific phonemes. Modern technology enabled us to check these theories by incorporating them into models of the speech recognition process. The first models failed to recognize speech successfully, showing that our theories were inadequate. As the years passed our ideas about speech recognition – and the models built to implement them – became more and more sophisticated. Yet, the results are still unsatisfactory. As matters stand today, we have, on the one hand, the human being, who can recognize speech with ease even under conditions of severe noise and distortion, and on the other hand, models of the speech recognition process which take into account all we know about human speech recognition and yet are able to deal only with fewer than a dozen or so words and only when spoken in isolation rather than connected text. It is only natural therefore that new and promising theories should be continuously proposed about various aspects of speech perception. One of these is the *motor* theory of speech perception. The motor theory proposes that, during speech recognition, we do not directly associate the sound qualities we perceive with linguistic units, the phonemes, words, etc., but that, instead, we first interpret our auditory percepts in terms of the articulatory movements needed to produce these sounds and, in a second stage, we recognize the language units by

association with these articulatory movements. A corollary of this theory is that an essential part of the process of learning to *recognize* speech is training in *producing* speech ourselves. The purpose of the experiment to be described in this paper was to observe how far being able to listen to our own voice and thereby getting a chance of associating the articulatory movements we make with the sounds produced by these movements makes learning to recognize speech easier. In this way it was hoped to learn more about the motor theory of speech perception.

In order to carry out the experiment, a method was needed for making naturally produced speech sounds sufficiently unlike normal speech that it could not be recognized without learning. At the same time, the sounds had to retain enough information about articulatory movement to make them learnable. This was achieved by modifying an eleven channel vocoder to compress the spectrum of speech in the ratio of roughly 3 to 1. The system is explained in Fig. 1. The 180 cps to 4500 cps wide speech spectrum is split into eleven

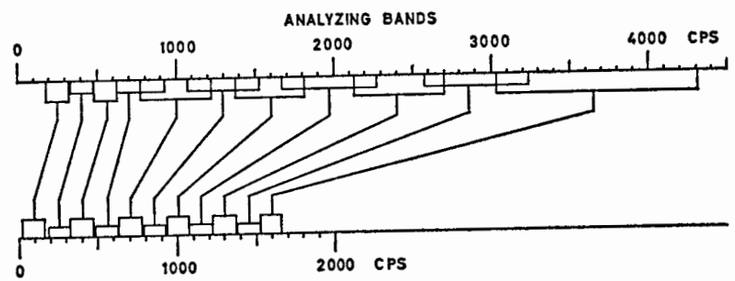

Fig. 1. Transposition of speech spectrum, as used in this experiment.

bands; the energy from each band controls the output of eleven synthesizing channels which, between them cover the frequency band of 50 cps to 1600 cps. The pitch of the speech input is also monitored and is used to control the frequency of the buzz source applied to the synthesizing filters. Again the pitch of the output, compared with that of the original speech input, is reduced in the ratio of about 3 to 1. The pitch of the output was reduced in order to have sufficient density of harmonics to excite the narrower band synthesizing filters. It will be noted that there are four analyzing filters to cover the first formant range and four to cover the second

formant range. The first formant filters are about 150 cps wide and the second formant filters about 450 cps wide, whilst the three highest filters are even wider. The speech input could be obtained either from a prerecorded magnetic tape or from a microphone. The output could be heard over earphones.

Ten senior high school students, five men and five women, were used in the experiment. Each session consisted of a twenty minute training period followed by a five minute test. The speech material used was taken from a library of 150 words. The words were pronounced in isolation, by one speaker, prerecorded on magnetic tape and rerandomized every time they were used. During the training period, the subjects listened to the processed words over earphones and, at the same time, they had a printed list of the words in front of them. In this way they could associate the sounds they heard with the words they represented. During the test period, a different list of words was used, the subjects still heard the processed speech, but they did not have the printed list and instead had to write down the words they could recognize.

The subjects were divided into two groups: the so-called "Listeners" and the "Speakers". During the training period the "listeners" could learn solely by associating the processed sounds of the one, prerecorded speaker with the words on their printed list. The "speakers", on the other hand, in addition, had a microphone. They first listened to the prerecorded voice and then had to repeat the word into their microphone. Their voice was processed by the identical device as the voice of the prerecorded speaker and they heard the processed version of their own voice over their earphones. The sounds from the earphones had a sufficiently high level to ensure that they masked any unprocessed version of their own voice which may have reached their ears directly, by either air or bone conduction. In this way they could learn to relate 1. the words shown on the printed list, 2. the associated articulatory movements they themselves made when they pronounced these words, and, 3. the processed sounds which they heard as a result of their articulatory movements. The five minute test which followed each training period was of course the same for "speakers" and for "listeners". The test lists consisted of 50 randomized words and the proportion of words recognized correctly was taken as a measure of the learning achieved. A comparison of the learning progress of the two groups was considered as an indication of the value, in recognizing speech,

of the articulatory associations available to the "speakers" only and thereby furnish an indication of the validity of the motor theory. The learning progress observed in 15 consecutive training periods by "speakers" and by "listeners" is shown in Fig. 2. No clear-cut

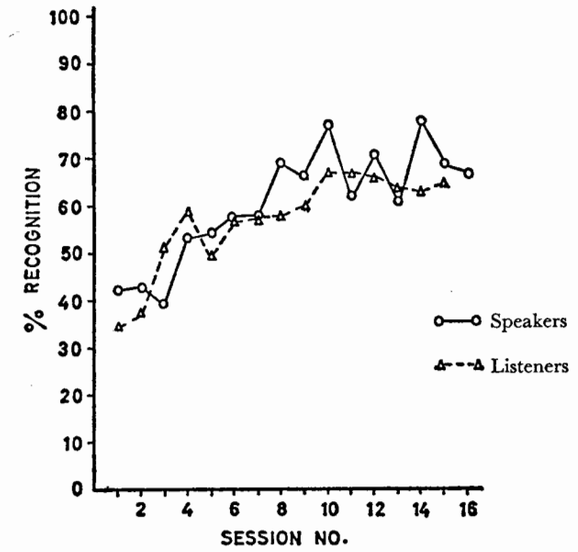

Fig. 2. Learning progress.

difference between the two groups is evident: both curves rise from roughly a 40% word recognition score to about 70%. The results therefore *do not* support the motor theory of speech perception. Just the same, a comparison of the speakers' and of the listeners' learning curves show certain dissimilarities that *may* indicate that the speakers' learning progress *was* influenced by the sound of their own articulations. There are, for example, the marked oscillations in the speakers' curve as compared with the relative smoothness of the listeners' curve. This may be the effect of the more variable sounds heard by the "speakers" as compared with the "listeners". The listeners always heard the same recording of only one speaker, whilst the speakers may well have varied their pronounciation as their learning progressed, producing a feed back type oscillation in their results. Also – perish the thought – some equipment failure cannot be disregarded. The test equipment was basically a vocoder and vocoders' pitch detectors are notorious for making errors under

certain combinations of speaker's voice, vowel quality and voice pitch. Whilst this could be controlled for the one prerecorded voice, pitch detector errors were observed for some of the other voices.

Further analysis of the recognition scores showed that almost three times as many errors were due to place-of-articulation than to manner-of-articulation confusions. This, of course, is quite consistent with the very rough quantization of the spectrum by the analyzing filters: the second formant region – so important for identifying the place of articulation of consonants – was divided into just four bands, each about 450 cps wide, so that only the really strong formant transitions were at all noticeable in the spectrum of the output. It may, in fact, have been more reasonable to have used a *formant* vocoder based system rather than the channel vocoder actually used. In this way, a much finer grain formant search at the analyzing end, would have allowed the compression of more formant information at the synthesizing end into the same narrow band as was used in the present experiment. By hindsight, this approach would have produced test signals that were still unlike normal speech yet richer in distinctive information about articulatory activity, making the learning of articulatory – auditory associations easier.

Was the output of the present system, in fact, distinctive enough to be learnable? On re-examination, the data showed definite evidence of the learnability of our nonspeechlike test sounds. The recognition scores obtained for certain key words after the first three learning periods were compared with the scores for the same words after the 9th, 10th and 11th learning periods. The confusions in recognizing word pairs such as for example *mirth* and *nurse* were examined. It was found that whilst at the start of the learning period confusions, were as high as 50% or more, by the end of the 11th learning period they were down to a very small value. And so on, for numerous other examples. The evident learnability of these processed "speech" signals is important not only theoretically but it also furnishes a guide to the promise of certain frequency-compression hearing aids. Unfortunately we have no time to discuss these hearing aids and the associated problems of relearning in this paper.

In conclusion then, it can be said that the tests have produced no firm evidence to support the motor theory of speech perception. The results have, however, shown some differences in the learning behavior of "listeners" and of "speakers", indicating that longer learning periods or perhaps more learnable signals might confirm

the motor theory after all. The results have also shown that some, at least, of the frequency transposed signals *were* learnable: an interesting result in itself. And finally, for those of you, who like myself, like the motor theory – because of the supporting evidence of psycho-acoustic results from Haskins and because of the unifying simplicity it promises for much of what we know about speech perception – there is a further point which I should perhaps have mentioned at the beginning of my talk: perhaps the kind of speech learning performed by adults – such as the subjects in my experiment – is different from that which takes place when a child learns speech.

Author's address: Dr. Peter B. Denes, Bell Telephone Laboratories Inc., *Murray Hill, N.J.* (USA)

Discussion

Jassem (Poznań): 1. I am not sure that the attempt has been made to construct a mechanical speech recognizer which possesses all the knowledge of language that has been collected by linguists.

2. Although Mr. *Denes* is not committing himself on this point, his paper appears to support the opposition against the motor theory of speech perception. There are many questions which would have to be answered before the theory can be unreservedly accepted. I will raise two:

(A) What is the delay time of the feedback system postulated by the theory? It might turn out to be so long that it might be an obstacle rather than a help in speech perception.

(B) What is the explanation of the fact that some speakers of a language (e.g. immigrants) have no difficulty in understanding its spoken form although they do not pronounce the language correctly?

Answer *Denes:* Mr. *Jassem* indicated that existing models of the human speech perception process could be more successful if only engineers would include all that is known about speech perception in the design of their automatic speech recognizers. Although it is true that in the past specialized engineering training, as well as knowledge of phonetics and linguistics was required for designing automatic speech recognizers, the availability of modern digital computers has changed this. Any phonetician or linguist, even if he has no knowledge of engineering, can put his theories on speech recognition to the test by computer simulation. All that is needed is a flow chart which specifies the sequence of logical operations which he considers are needed to implement the recognition process. If Mr. *Jassem* thinks that engineers have not utilized all that linguists and phoneticians know about speech recognition, I would ask him to specify the principles which we ignored. Any qualified computer programmer will be able to include his ideas into a suitable program and put them to the test.

Tillmann (Bonn): I want to ask you: Did you investigate whether the best of the trained subjects were able to identify even quite new not trained words presented to them after frequency compression?

Harms (Lawrence): 1. In preparing a learning program for phonetic transcription, students were observed to repeat the word they heard before they attempted to transcribe

it. How might this be interpreted within the framework of the Motor Theory of speech perception?

2. What is a next experiment to follow the one you have completed?

Fry (London): I should like to make one general remark and also one specific comment on the experiments reported by Mr. *Denes*. The general point is one already mentioned by Mr. *Jassem*. Let me put it in this way: every individual spends a good deal of time acting as a speaker and a good deal as a listener. We are bound to assume that much of the brain mechanism is common to both processes. It would be entirely against what we know of the economy of the human being to imagine that there are completely separate mechanisms for the generation and for the reception of speech. On the other hand, there *are* differences between reception and generation and there are no doubt some mechanisms that are specific to the one or the other. For this reason I suggest we should not refer to "the motor theory of speech perception", nor even to "a motor theory" because speech can be perceived in a number of ways. Perception of speech at any time will be the result of a number of different factors, and what we are trying to find out is what weight should be given to the effect of motor memories in given circumstances. I have said before, for example, that I believe this factor has considerable weight in the perception of stress patterns.

The specific point concerns the task that subjects were asked to do in these experiments. It seems to me that the severity of the task in the case of the "speakers" has been rather under-rated. They are being asked not simply to learn associations between speech movements and sounds; they have first to dissolve existing associations which link articulatory movements with normal sounds before they can establish the new ones and this is clearly a very difficult thing to do. It would probably take quite a long time to achieve this and it is not surprising that the effect should not show up after the relatively short training period allowed in the experiments. It might be worthwhile to get both "speakers" and "listeners" up to some fairly high level of performance and then see how rapidly both groups could learn a fresh set of words.

Answer *Denes:* I agree completely that the motor theory can only represent *a* process of speech perception, rather than *the* process. Like with so many other aspects of the human speech communication process, or other human activities for that matter, the human being utilizes a variety of ways for achieving his aims. In speech perception, the motor theory probably accounts for only one for several ways in which perception is established.

Fourcin (London): Although the average performance for each of the two groups of subjects, those who spoke and listened and those who only listened, is almost the same, the difference in detailed shape of the two curves could be of importance. The speakers may, because of their prior practice, learn at a greater rate than the listeners during an experiment. In between sessions however they may at first continue to learn from their accumulated experience and then, if the interval is of sufficient length, start to forget owing to the disturbing influence of normal speech. Thus if the temporal spacing of experiments is short-long, the speakers' zig-zag performance curve could be expected. If these assumptions are correct, the zig-zag curve would be modified by having more experimental sessions in a working week. Then, if the inhibiting effects of fatigue are not pronounced, the speakers would have a greater average score than the listeners.

From this it seems possible that the basic experiment may have been more successful than at first appeared; it is not, however, of crucial consequence to the motor hypothesis, too many unknown factors are involved.