Relation between Speech Production and Speech Perception

L.A. Chistovich
Leningrad, USSR

1. Introduction

The problem of relations between auditory representation of speech unit and the 'goal' in the program for this unit production is highly important for both speech production and speech perception theories. Auditory control of timing in execution of motor program is another aspect of the problem. It was supposed that the acoustical events arising at the onset of speech sound production might trigger, after prescribed delay, the execution of the next speech unit (the 'chain' model of production).

Neurophysiological studies of the central auditory system suggest the extraction of two kinds of information from the peripheral auditory pattern - one, most appropriate for timing control (short 'phasic' responses to rapid spectral and amplitude changes) and the other one, more appropriate for specification of the goals (selective responses to specific spectrum shapes, direction of spectral peak transition and so on). It is important to note that auditory neurons seem to have only a primitive memory: they can integrate, with some time constants, the incoming excitations and inhibitions and they can become temporarily blocked after firing. The time window of processing appeared to be different for different neurons but it did not exceed 200 ms.

The aim of this paper is to review some experiments where external speech stimuli were used to control speech production. In speech-by-speech synchronization experiments the subjects produce the prescribed response, only the timing might be controlled by the stimulus. Experiments on mimicking concern the goals formation. Both the goals formation and the timing are involved in shadowing.

2. Speech-by-speech synchronization

The subject can synchronize the production (response) with periodically presented stimuli (clicks, tone pulses) and make stimulus and response overlap in time. It was speculated that if speech stimuli were to be used for synchronization, the speech execution mechanism might mistake the marker of the speech sound onset in the stimulus for the marker of the corresponding onset in the response. It was found: Chistovich et al. (1972) that the interval between V_1 onset in VCV-stimulus and V_2 onset in VCV-response was really

more stable than the interval between V_1 and V_2 onsets in the response. The variability of the last interval (between V_1 and V_2 in response) appeared to be much higher in these conditions than in both free and synchronized by clicks VCV productions.

3. Mimicking of isolated sounds

Experiments on mimicking the loudness of fricative sound (Malinnikova, 1968), vowel duration (Chistovich et al., 1966) and tone pitch (Lublinskaja, 1968) were at first aimed to test whether the subject's goal was to match the response with the stimulus. Clear negative results were obtained: the subjects were very good in preserving in responses the orderly relations among stimuli but they did not reproduce the absolute values of the stimuli. (Mimicking of pitch by musically trained subjects was an exception). This points to wired up scales relating the auditory system outputs to motor control parameters. The efforts were aimed at finding out whether these scales are fixed or adaptive and when the last alternative appeared to be true, to study the variables controlling the scale adaptation (Malinnikova, 1971).

Experiments on mimicking synthetic vowel by subjects with different sizes of their vocal tracts (males, females, children of different ages) have shown that the subjects preserve the orderly relations among stimuli in formant frequencies space but they do not match the spectrum of the response with that of the stimulus (Kent, 1978; Kent et al., 1979). There are indications that vowel mimicking is an innate behavior. The important problem is to find out whether continuous or discrete scales relate responses to stimuli in vowel mimicking. Clustering of responses predicted by categorization has been observed (Chistovich et al., 1966; Kent, 1973; Kent, 1978) but far more extensive data are needed for a reliable conclusion.

4. Shadowing

The ability of subjects to rapidly imitate (shadow) natural and synthetic speech is a well documented fact. The data on shadowing stop consonants in VCV stimuli are best suited to discuss the implications of the effect and the problems involved in its analysis. Identification experiments have shown that although some information about consonant identity is conveyed by closure transition and initial part of closure, the subjects rely in phoneme decision on the events following the release of closure. In shadowing VCV the subjects start the consonant production before the release of closure in the stimulus (Kozhevnikov et al., 1965; Porter et al., 1980). That means that auditory information corresponding to closure transition is transformed into motor representation (goal or the set of goals) and could be stored in this form till new auditory data arrive. It was found that consonant response might begin with erroneous articulation, which could be corrected in the course of production. It was tempting to speculate (Kozhevnikov et al., 1965)

that response modifications in shadowing reflect the temporal process of phonetic interpretation. I shall present some arguments against this view and in favour of the idea that the execution of the motor program observed in shadowing and the formation of this program might appear to be two parallel processes controlled by a different kind of auditory information. The latencies of shadowing are equal to simple reaction time. It was found that the actual signal eliciting the response in experiments on simple reaction time to tone is not the tone but the event of onset (presumably on-response). Substitution of the stimulus with the tone of far different frequency results in the same response with just the same latency (Chistovich, 1956). If shadowing resonse to vowel is also triggered by the onset of the stimulus, then by cutting out the late parts of the vowel we might influence the quality but not the latency of the response. The experiments on shadowing the natural whole and truncated vowels confirmed these expectations. The critical stimulus duration determining the initial part of response appeared to be between 50 and 100 ms (Chistovich et al., 1962).

The experiments on shadowing synthetic /ao/, /aæ/ and ai/stimuli with long and variable /a/ duration (Porter et al., 1980) have shown that subjects start correct response to second vowel with a latency of 150 ms from the onset of the formant transitions. The same or a little longer latencies were observed in simple reaction time situation: subjects had to respond by /ao/ to all three kinds of stimuli. This also suggests that the same events trigger the response execution in both tasks.

5. Mimicking of simple sequencies.

Comparison of mimicking response to isolated stimulus with the response to the same stimulus in contest seems to be a good approach to study contextual rules. Pronounced contrast effect has been observed in formant patterns of the second vowel produced in mimicking VV-stimuli with different first vowels (Kent, 1974). It was also observed in vowel durations produced in mimicking VV with different durations of the vowels in the stimulus (Zhukov, 1971). Pitch contrast effect was studied on musically trained subjects, who were instructed to listen to a tone pair and precisely reproduce both stimuli. The subjects followed the instruction when the frequency difference between stimuli was large. When it was small, they made one response higher and the other one lower than the corresponding stimulus. It seems that the subjects tried to preserve the average to the pair pitch and to increase the difference between components of the pair (Lublinskaja, 1970). It is clear that this kind of processing is not compatible with the facts concerning the auditory system. True memory and the ability to read out and modify the previously recorded item are necessary.

6. Concluding remarks

It is obvious that the brain must possess some 'language' to translate the auditory information into information to the motor system. The data on shadowing suggest that the translation occurs with short delay and does not require long auditory memory. The results on mimicking suggest that this audio-motor 'language' is at least partly innate. It is tempting to speculate that several phonetical effects and regularities reflect in fact the structure and the rules of this 'language' and could be found under close examination in various perceptual-motor skills.

References

Chistovich, L.A. (1956). Comparison of conditioned motor reactions based on shock and verbal reinforcements. *Fiziol. zh. USSR,* **43**, 572-580.

Chistovich, L.A., Fant G., de Serpa-Leitao, A., Tjernlund, P. (1966). Mimicking of synthetic vowels. *Quart. Progr. Status Rep. Speech Transm. Lab. Roy. Inst. Technol. Stockholm* **2**, 1-18.

Chistovich, L.A., Klaas Ju.A. (1962). Toward analysis of latency of 'voluntary' reaction to sound. *Fiziol. zh. USSR,* **48**, 899-906.

Chistovich L.A., Lissenko, D.M., Fedorova, N.A. (1972). Speech production control under synchronization with periodically presented speech stimulus. In: *Sensornye sistemy,* **2**, Leningrad, 56-85.

Chistovich, L.A., Zhukova, M.G., Malinnikova T.G., Kozhevnikov, V.A., Borozdin, A.N. (1966). Mimicking and perception of isolated vowels. In: *Mechanizmy recheobrazovania i vospriatia zlozhnych zvukov.* Leningrad, 128-157.

Kent, R.D. (1973). The imitation of synthetic vowels and some implications for speech memory. *Phonetica,* **28**, 1-25.

Kent, R.D. (1974). Auditory-motor formant tracking: a study of speech imitation. *J. Speech Hear. Res.,* **17**, 203-222.

Kent R.D. (1978). Imitation of synthesized vowels by preschool children. *J. Acoust. Soc. Amer.,* **63**, 1193-1198.

Kent, R.D., Forner, L.L. (1979). Developmental study of vowel formant frequencies in an imitation task. *J. Acoust. Soc. Amer.,* **65**, 208-217.

Kozhevnikov, V.A., Chistovich, L.A. (1965). Speech: Articulation and perception (Translated from Russian). *Report* **30**, 543. Washington, D.C.: Joint Publication Research Service.

Lublinskaja, V.V. (1968). Mimicking of pitch. *Ztschr. für Phonetik* **21**, 129-134.

Lublinskaja, V.V. (1970). Mimicking of pitch interval in sequences of two tonal stimuli. In: *Upravlenie dvizheniami.* Leningrad, 110-117.

Malinnikova, T.G. (1968). Mimicking of loudness of synthetic fricative consonants. *Ztschr. fur Phonetik,* **21**, 135-139.

Malinnikova, T.G. (1971). Relations between response intensity and the intensity range of stimuli in consonant mimicking. In: *Sensornye sistemy,* **2**, Leningrad, 99-110.

Porter, R.J., Castellanos, F.X. (1980). Speech production measures of speech perception: Rapid shadowing of VCV syllables. *J. Acoust. Soc. Amer.,* **67**, 1349-1356.

Porter, R.J., Lubker, J.F. (1980). Rapid reproduction of vowel-vowel sequences: Evidence for a fast and direct acoustic-motor linkage in speech. *J. Speech Hear. Res.* **23**, 593-602.

Zhukov, S.J. (1971). About auditory segmentation of /iu/. In: *Sensornye sistemy,* **2**, Leningrad, 71-82.