THE INITIAL TIME SPAN OF AUDITORY PROCESSING USED FOR SPEAKER ATTRIBUTION OF THE SPEECH SIGNAL.

V.V. Lublinskaja
Pavlov Institute of Physiology, Saint-Petersburg
Tel. +7 812 529 09 58, Fax: +7 812 218 05 01, E-mail:chi@physiology.spb.su
Ch. Sappok
Institute of Slavonic Studies, Ruhr Universität Bochum
Tel. +49 234 700 6664, Fax: +49 234 7094 337, E-mail: sappokc@slf.ruhr-uni-bochum.de

ABSTRACT

Research on the temporal organisation of speech perception is focussed mostly on the linguistic categories of the input. What is the role of non-grammatical categories for this processes? What kind of mechanisms integrate both kinds of features within the online process of perception? Individual voice qualities and the position of the sentence within the text were chosen to test the time interval where decisions as to speaker belongingness are made. The results favour a model with a relatively fixed time span within which a familiar voice or a deviation from an inherent context expectancy are detected.

1. TASK.

Whereas research on the time organisation of speech perception is focused on the processes of phoneme identification or lexical access, few is known about the timing of the auditory processing of so-called extralinguistic factors as, for example, speakers' voice which is known to play an important role in the recognition of speech. A first attempt to procede in this direction was made by Lublinskaja , Sappok 1996. The task of our present work is to investigate the temporal side of processing speech signals when a listener has to ascribe them to some familiar speaker. Being involved in the process of modelling the discourse situation he has to trace a target voice within sequences of sentences spoken by different speakers.

Two questions have to be answered: (1) How long is the initial time span subjects need to ascribe sentences to a target voice? 2) What is the nature of this interval: does it depend on the specific features of the acoustic events? Or is it a standard rate of scanning the results of auditory input as represented in memory? Hypotheses on this field have mostly been discussed in connection with phoneme identification (Chistovich 1984, Massarro 1972) or with lexical access (Marslen-Wilson 1985) paying little attention to online processing of speakers´ voice characteristics.

2. STIMULI.

A set of sentences from the Acoustic Data Base of the Saint-Petersburg University was used as speech material [2]. Two main tests preceded by two preparatory tests (cf. below) were prepared: The first contained a sequence of 28 sentences spoken by two alternating voices (female and male). In the second test 31 sentences where spoken by four speakers: two female and two male. In both tests one of the female voices was chosen to be the familiar voice being the object of training procedures, the other voices being unfamiliar. The familiar voice predominated over the others: they occurred three times as often as compared with the unfamiliar ones. The pitch characteristics of stimuli were: for the familiar female speaker - F0=166 Hz (mean level), and 357 Hz (max); for the unfamiliar female speaker - F0=240 Hz (mean), and 450 (max); for the first male speaker - F0=120 Hz (mean), and 200 Hz (max); for the second male speaker - F0=100 Hz (mean), and 120 Hz (max).

The sentences varied by duration from 0.3 to 3.4 sec. being presented in random order.

In some cases they remained thematically connected because of their stemming from a semantically coherent text. Nine experts (phoneticians from the Saint-Petersburg University) had to listen to the sentences in isolation and were asked to judge whether a continuation of this stretch of communication is expected or not. If the answer as to the continuation was positive subjects were asked to decide whether it was expected as coming from the same speaker or not. On the base of these attributes all sentences were marked by one of two parameters - contextual neutral vs. containing inherent cues predicting continuation. In the letter case stimuli were separated: those which coincided with the above mentioned expectancies and those which did not coincide.

3. PROCEDURE.

An IBM PC with audio equipment (12 Byte resolution, 20 kHz sampling) and the software presented by [2] were used to prepare the stimuli and to realise the following experiments. The experimental procedure consists of four sessions performed within one day. During two main sessions Test 1 and than Test 2 were presented. Both of them were organised so that in the first part subjects were trained to be acquainted with the voice of one speaker

considered as "familiar" or "target" in the subsequent series of experimental tasks. Each subject listened to one sentence (1.5 sec) spoken by this speaker repeatedly till he became sure to be able to remember this speaker's voice. Than the sequence of sentences was presented in random order. The task was to decide for every sentence whether it was pronounced by the familiar voice or by some other voice. The answer had to be given as quickly as possible pushing on one of two knobs on a small key-board. The subjects' responses and reaction intervals measured from the onset of the stimulus were recorded.

The main sessions were preceded by two training sessions where subjects had to become acquainted with the procedure and with the task of the experiment. The first of them was similar to the main tests but consisted of different sentences which could be repeated as many times as a subject wanted.

In the second additional experiment the subjects had to react to the appearance of sweep tones of equal frequency area (1000-1500 Hz) presented successively at random time intervals. The task of this session was to evaluate the time of simple sensor-motor reactions (henceforth SMRT) of every subject.

During all sessions subjects were seated in a noise-proof boot and were listening to stimuli through earphones with comfortable level. 22 Russian adults without hearing problems participated in the experiments.

4. ANALYSIS OF THE RESPONSES

The difference between response time measured in the main experiments (RT) and the SMRT was used as a value which is supposed to be close to the time of processing of signals on the stages before a motor realisation of decisions. We call it conditioned time processing (henceforth TP). The responses were taken into account only for those subjects whose SMRT was less than 0.225 s.(maximum time reaction in of subjects participated in experiments) The responses of 20 subjects were chosen for analysis: TR and the errors of the answers.

5. RESULTS

The overall analysis of all answers given in the two main tests shows the following results.

5.1. The frequency of correct decisions concerning the familiarity of voices was 89 % in the first test and in 83% in the second test. Decisions concerning unfamiliar voices were more correct: 97% and 95% in the first and the second test, respectively.

As one can see the behaviour of the subjects does not depend on the circumstance whether the familiar voice has to be recognised among two or more voices. But it is worth to note that features of unfamiliar voices (cf. The description of stimuli above) differ considerably. Most errors occurred with sentences spoken by the second female voice.

5.2. The values of TR for all stimuli create asymmetric distributions with a prominent maximum in the time interval between 0.4 and 0.5 sec. This result was valid for both tests. It can be seen in Fig. 1, where both distributions are plotted.

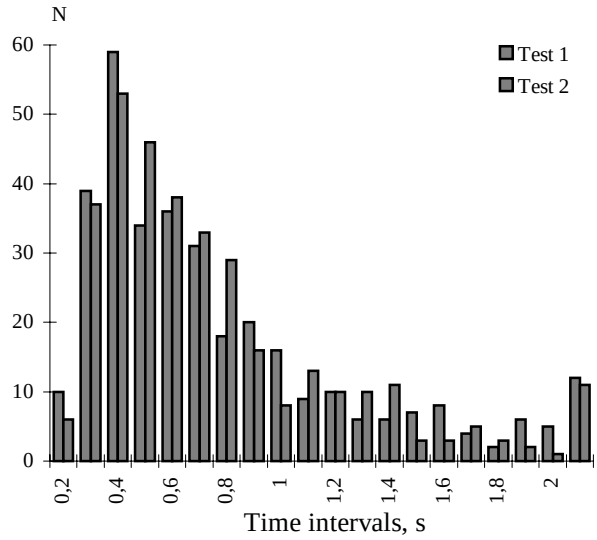

Fig. 1. Distribution of response delay (TP) for Test 1 and Test 2.

The correspondence of decisions in both tests was proved also by statistical comparison between the time delay of the answers for identical sentences (there were 18 of them) included in the both tests. (We used T-test with equal variance).

5.3. The most important information obtained concerns the relation between response delay and the duration of sentences. It is illustrated in Fig. 2.

As can be seen response delay slightly increases with the duration increment of the stimulus sentence. But the increase ratio is different for short and long stimuli. For short stimuli (up to 0.75 s approximately) a strong correlation between TR and sentences duration was observed (r=0.79). For longer stimuli TR was smaller than sentence duration and the correlation between them was week (r=0.3).

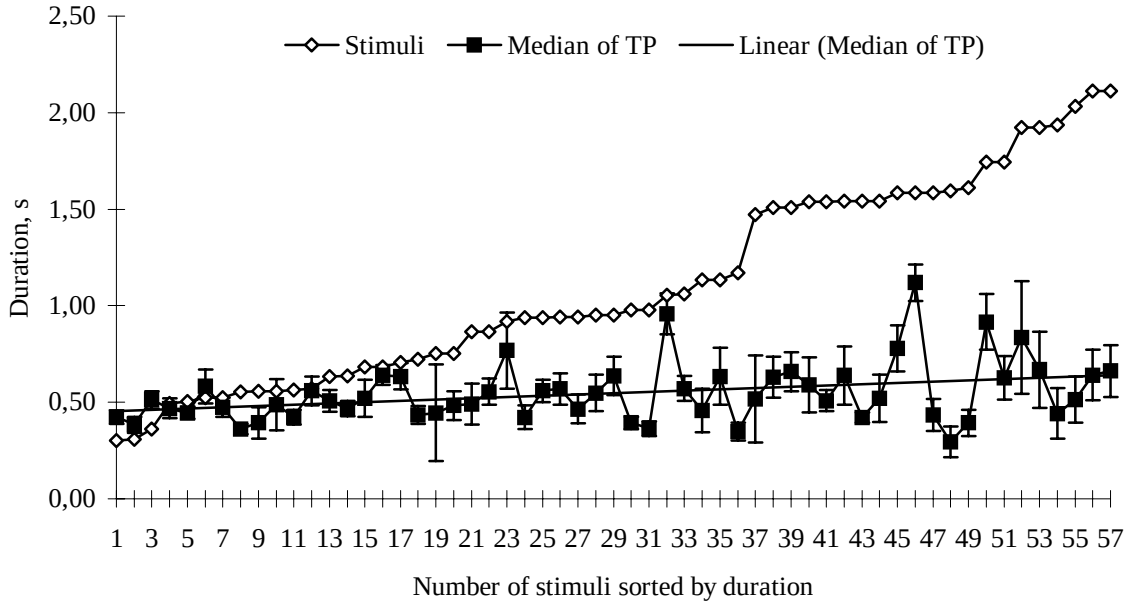

Fig. 2. Response delay (TR) in relation to the duration of stimuli.

5.4. A remarkable effect of the inherent expectancy of the sentences as described above was observed: the average TR for stimuli with expected continuation was 0.1 sec shorter than for contextual neutral ones. The erroneous answers and the TRs reflect an influence of this contextual markedness: there were 13% of errors in cases where no context is expected, 7% in expected cases when the stimuli coincided with expectancy, and 15% in cases without this coincidence. Mean (median) values of TR were: 0.73 (0.54), 0.64 (0.47), and 0.59 (0.51) corresponding to the three cases mentioned. The differences were statistically proved (T-test) with the confidence level 0.05.

6. CONCLUSION.

The concentration of the TR within a narrow interval of values and a weak correlation between the duration of long sentences can be interpreted as evidence that some initial critical interval of processing of input information concerning voice quality of the speaker exists.

At the same time a deciding role of inherent features of the acoustical event seems impossible to be rejected. The strong correlation of time delay with the duration of short stimuli give evidence for the synchronisation of processing time with the offset of signals when the last one is in the vicinity of this critical interval.

The question arises whether some other feature of the acoustic event is used as the place were the processing of auditory input ends and the choice is being made. The end of the stressed syllable can be assumed as the most probable candidate. An attempt was made to test this assumption comparing the distributions of TR and the offset time of stressed syllables in sentences. Fig. 3 A, B shows the results for the first and for the second tests separately. As can be seen the envelopes of distribution for both dimensions does not coincide completely. This gives reason to weaken the evidence of the proposed assumptions.

To explain the deviating portions of the overall picture, especially the crucial relation between TR and sentence duration, it seems necessary to go into the details of the segmentation of the input speech flow. But this remains the topic of future work.

Another problem remains to be considered: Is the processing of voice quality unconnected with linguistic knowledge at all?

We tried to find an answer to this question by carrying out the same experimental procedure with subjects whose native language is German and who didn´t learn Russian as a second language. 15 students from the Ruhr University Bochum participated in the experiments.

The results were the following: The most remarkable difference between Russian and German speaking subjects shows up in the answers to the familiar voice. Correct answers were given by 85% of the Russian subjects and by 75% of the German subjects. The time delay (median of TP) was 0.56 sec and 0.89 sec., respectively. The difference in the case of unfamiliar voices was less prominent: correct answers were 95% and 97% and the

median of TP was 0.43 and 0.53, respectively. (The difference statistically is not significant at the level 0,05).

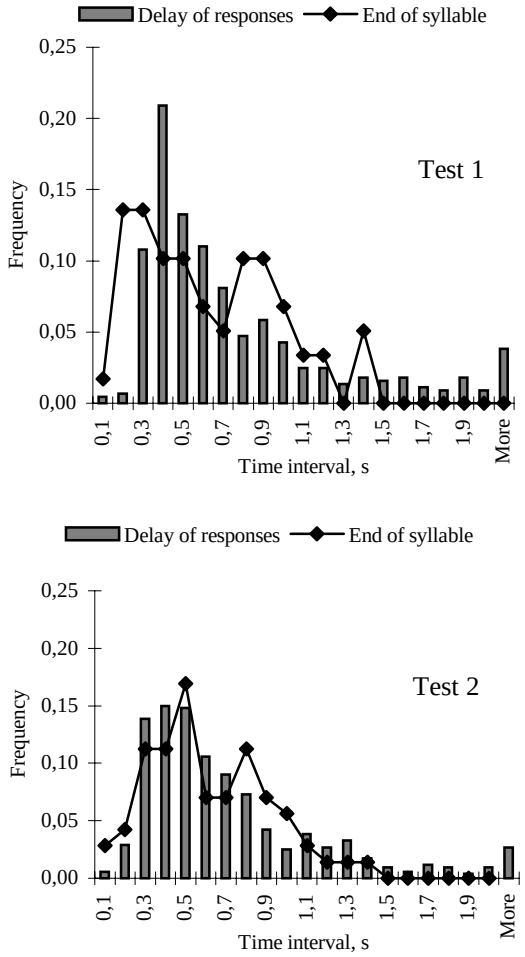

Fig. 3. Comparison of the distributions of response delay (TP) with the ending of stressed syllables.

The results of the experiments where the decision concerning speaker belongingness has to be made without understanding what is said allow to suppose that grammatical knowledge plays a role in the process of speaker attribution. Probably memory is the deciding factor. That means that information concerning voice is not stored in memory in the form of isolated acoustic features. These seem to be accompanied by linguistic attributes.

References

[1] V. Lublinskaja, Ch. Sappok, *Speaker attribution of successive utterances: The role of discontinuities in voice characteristics and prosody*. Speech Communication, 1996, 145-159.

[2] L.V. Bondarko et al., *Fond zvukovykh edinic russkoj rechi*. Bjulleten´ foneticheskogo fonda russkogo jazyka, Prilozhenie Nr. 3, S. Peterburg - Bochum, 1993.

[3] L. Chistovich, *Auditory processing of speech.* Language and Speech, 1984, 23, 67-73.

[4] D.W. Massaro, *Perceptual images, processing time, and perceptual units in auditory perception*. Psychol. Rev., 1972, 79, 124-145.

[5] W.D. Marslen-Wilson, *Speech shadowing and speech comprehension.* Speech Communication, 4, 1985, 55-73.

[6] M. Knipschild and Ch. Sappok, *Akustische Zeichnenverarbeitung durch SONA und VERSTEU*. Fortshritte der Akustik - DAGA, 1991, 1045-1048.