EVALUATION OF SYNTHETIC SPEECH QUALITY: A COMPARATIVE STUDY OF SEVERAL COMPUTER-BASED SPEECH SYNTHESIZERS*

Albinas Bagdonas

Vilniaus universiteto Socialinio darbo katedros
ir Specialiosios psichologijos laboratorijos
vedėjas, gamtos mokslų daktaras
Universiteto g. 3, 903, LT-2734 Vilnius
Tel. 68 72 53, faks. 68 72 90
El. paštas: Albinas.Bagdonas@fsf.vu.lt

Feliksas Laugalys

Vilniaus universiteto Bendrosios ir pedagoginės
katedros docentas, gamtos mokslų daktaras
Didlaukio g. 47, LT-2057 Vilnius
Tel. 76 25 72
El. paštas: Feliksas.Laugalys@fsf,vu,lt

This paper examines some versions of Lithuanian and Russian synthetic speech intelligibility and Lithuanian, Russian, Hungarian and Italian synthetic speech acceptability. The speech of both Russian and Lithuanian speaker is more intelligible than Russian or Lithuanian synthesis. Previous version of Russian synthesis is worse than Lithuanian and improved Russian synthesis (IRS). Study of characteristics of IRS sounds shows two opposite tendencies – according to the general quantity of mistake reduction this version is tending towards the natural speech, but according to the homogeneity of mistakes, it moves away. As the first tendency is clearly dominant, the general resultant in the new version shows a tend to improve.

Correlation between intelligibility and acceptability of IRS deals possibility of small progress towards the natural speech. The IRS is more acceptable to subjects than previous version. The old synthesis is viewed as a rather decent instance of a robot's speech, while the IRS – as a poor variant of human speech.

Acceptability studies showed natural speech more enjoyed by Hungarian listeners and more critical by Italian. All versions of synthetic speech were judged as less acceptable than natural but after improvement most of listeners changed their mind.

1. Introduction

1.1. Speech Perception and Computerized Speech Synthesis

Perception of speech (spoken text) is an exceptionally complex cognitive process which includes a number of sub-processes, such as sound analysis, speech units of different levels identification, memory, comprehension, individual emotional appraisal, etc. In the course of human development each language has formed its own principles of speech generation and perception. Therefore it is understandable that creators of various acoustic speech-transferring systems se-

*This investigation was granted by DG of the Comission / commission of the European, project Copernicus, Digibook – 806 (Digitized Speech Processing for Efficient Distribution of Texts).

ek to achieve these natural principles. It especially concerns speech synthesis. This is why research work continually aims at revealing still more advanced methods of speech synthesis (Blenkhorn, 1995; Rahim, 1994), seeking at the same time to distinguish and realize in synthesizers as many text parameters as possible (Werner and Keller, 1994; Marcus and Syrdal, 1995).

Two aspects can be distinguished in the verbal communication process: The speech-generation (human voice and speech producing system, some acousticspeech-reproducing system) and the speech perception (listener). In this context are two groups of methods for speech quality evaluation. The first group includes evaluation of physical speech parameters. All this usually constitutes the content of phonetics, a separate branch of knowledge (Keller, 1995). Parameters of generated phonetic units (frequency-amplitude structure, duration, reciprocities, etc.) can be easily evaluated by means of physical measurements.

However physical parameters of voice and pronounced phonetic units, as they are, cannot be absolute indices of speech quality, though the speech quality depends on them directly. Psychological impact of speech depends also on the recipient's ability to differentiate, integrate, interpret and comprehend sounds and units of speech. No wonder that creators of various acoustic systems (including the speech synthesizers), realy not only upon definite technical measuring devices when evaluating sounds generated by the said systems, but also check how they are perceived by listeners. Theoretically (and practically) it is possible to evaluate subjectively all parameters of generated speech, i. e. its intelligibility, pleasantness, level of "contamination" by noises, force, timbre, tempo, etc. (Preminger and Vantsel, 1995).

The phenomenon of speech can be characterized by a multitude of parameters. Consequently, speech perception is a complex process and, therefore methods used in speech quality investigations reflect this complexity of perception. Subjective psychological scales are created, speech fragments are compared, opinions given by subjects are interpreted, speeches generated by two speech generating systems are compared, (e. g. the natural speech and the speech generated by a synthesizer, speeches generated by two synthesizers).

The following four closely interrelated aspects of speech perception (and of it's quality evaluation at the same time) we would like to discuss: 1) Speech intelligibility; 2) Speech comprehension; 3) Generated speech requirements for cognitive processes (attention, memory, reasoning); 4) Speech acceptability.

1.2. Speech Intelligibility

Speech intelligibility is the most used and qualitatively as well as quantitatively expressible parameter of speech quality. The speech intelligibility can be evaluated:

a) directly by means of creating subjective scales, (e. g. evaluating voice clarity by 7-point, 10-point or 100-point scales); b) indirectly according to the number of correctly recognized phonetic units (speech sounds, words, sentences or phrases); c) by means of comparing (differentiating) two or more fragments of generated speech, or by ranking of a set of speech fragments.

Evaluation of speech intelligibility is the most common aspect of psychological investigation of speech quality (Kajinder and Allen, 1993; Preminger and Vantasell, 1995; Hoce and Pavlovic, 1994). Speech intelligibility perception is especially sensitive to a variety of factors. Therefore measurements of intelligibility (especially, by number of correct recognitions) help to evaluate the quality of speech generated by different sources, to determine various influencing factors: noise or speech transmission line peculiarities (Koul and Allen, 1993; Santon, Marchioni and Susini, 1994; Payton, Uchanski and Braida, 1994), the influence made by experience in using a synthesizer (Rounsefell, Zucker and Roberts, 1993; McNaughton, Fallon, Tod and Weiner, 1994), the speech fragment duration (Venkatagiri, 1994), the quality of cochlear implants (Osberyer and Sam, 1993), etc.

1.3. Speech Comprehension

Speech Comprehension measurements are essentially similar to speech intelligibility measurements. However in this case speech units of greater size – texts and their fragments – are used (Dufy and Pisoni , 1992). Though many other factors determine speech comprehension, it is closely related to speech intelligibility (Paris, Gilson, Thomas and Silver, 1995). Again, measurements of this type can be used to evaluate the influence made by various factors (Higginbotham, Drazek, Kowarsky and Scally, 1994; Higginbotham, Scally and Kowarsky, 1995). Speech comprehension investigations have been rather exhaustively reviewed by Ralston, Pisoni and Mullennix (1995) in their work.

1.4. Requirements for Cognitive Processes

Research of cognitive process load during the speech perception enables to evaluate various speech aspects. However the efficiency of this method and its sensitivity to various factors is not great, as it is necessary to take into consideration a multitude of possible factors when investigating the cognitive process load. For example, requirements for attention or memory can be increased not only because of a decrease in speech intelligibility but also because of other text parameters, state of the subject, etc. This method can be applied in two aspects:

a) When trying to evaluate the main speech parameters (intelligibility, comprehension and acceptability);

b) As an independent speech quality indicator (the speech synthesizer that requires less strain of cognitive processes is better in quality than the synthesizer that requires more cognitive efforts). Often in this way interaction of several of variables, e. g. synthetic and natural speech, age of subjects (old and young) and efficiency of memorizing is investigated (Smither, 1993; Humes, Nelson, Pisoni, Lively, 1993).

1.5. Speech Acceptability

If speech intelligibility characterizes chiefly the quality o voice and articulation, and speech comprehension reveals principally semantic characteristics of the text, speech acceptability is in the most part related to the emotional tone of speech (pleasantness of voice and articulation sound, speech organization level, logical structure of speech, etc.). Speech acceptability evaluation methods can be applied in the investigation of both short and long speech units, traditional or structured talking books. In the speech evaluations of this type a major role belongs to personality characteristics of the recipient: Gender, age, education, interests, etc. (Pavlovic, Rosi and Espesser, 1990; Tucker, 1991). Gorenflo et al. (1994) revealed that the quality of synthesis of used synthetic speech determines the attitudes towards disabled persons who use the speech in their communication; better quality of synthesis encourages more favorable attitudes towards its users. Attempts are made to increase the acceptability of synthetic speech by means of enriching it with emotional tone (Murray and Arnott, 1993, 1995).

Acceptability of speech, talking texts and books can be evaluated by means of the following threc basic methods: a) By comparing of pairs of excerpts of recorded or natural voices (e. g. more pleasant and less pleasant voice); b) Subjcctive scales (e. g. evaluating pleasantness of voices by 1–100 points scale); c) ranking of voices, talking texts or books (e. g. several different speeches are to be arranged in succession according to their pleasantness or naturalness).

1.6. Purposes of Present Research

The main purpose of present investigation was the comparative evaluation of the intelligibility and acceptability of the tape recorded specches of: 1) two Russian synthesizers (RS – the first version of Russian Synthesizer, IRS – Improved Russian Synthesizer); 2) Lithuanian Synthesizer (LS); 3) Lituanian / Lithuanian and Russian Speakers (LSP and RSP); 4) Russian talking books for blinds. By doing the project DIGIBOOK-806 we had possibility to compare our results with the results of other participants of this project (results of evaluation of Italian and Hungarian speech synthesizers and talking books).

2. Method

2.1. Speech Material, Subjects, and Design

Three kinds of tape recorded speech material were used. The first speech sub-test consisted of complete collection of Russian and Lithuanian letters (their acoustical equivalents). Such choice was predetermined by the goal of whole project – creation of speech synthesizer of high quality. For blind users of speech synthesizers it is important acoustic control of separate characters appearing on the computer screen. So intelligibility of synthesized phonemes is very important parameter of speech synthesizer.

The each word subtest consisted of 30 words which were randomly selected and mixed from the frcquency vocabulary: 10 words of high frequcncy, 10 words – of middle frequency and the last 10 words – of low frequency (Grumadienė, Žilinskienė, 1997). In the same way were prepared Russian word subtests.

The third kind of subtests consisted of 30 short (5–7 word) sentences. Sentences were created using one word in it of the same three frequencies as in word subtest.

The acceptability of synthetic speech was evaluated by mean of *Speech Synthesizer Appraisal Fonn – Questionnaire*, which consisted of 61 item – open and closed questions concerning listened speech unit acceptability, quality, experience in using speech synthesizers or talking books, possibilities of implementation of speech synthesis, as well as some personal characteristics of subjects. Some questions were presented in the form of 10-point scale. Because testing was conducted in normal environmental conditions for listening, we did not use separate evaluation of signal and background quality. We don't use too the explicitly identified anchors which other authors sometimes presented to the listeners as a frame of reference, but asked listener's to concentrate on their attitudes and emotional feelings and their changes during listening. At the beginning of acceptability evaluation a small listening probes were made to detect the level in listener adaptation. Some training and calibration and monitoring procedures were made too.

Subjects (listeners). One of the most important aspects, which arises during speech synthesis intelligibility and acceptability evaluation is a number of listeners, necessary for statistically reliable results. The IEEE Recommended Practice for Speech Quality Measurements (1969) recommends 6–10 trained listeners or at least 50 untrained (naive) subjects. So, training and calibration can considerably reduce the number of listeners needed because it results in decreasing of response variability. I n real life these procedures are prolonged (sometimes several weeks is needed), difficult for listeners and expensive. We solved this problem using different groups of listeners – subjects visually impaired group was trained with speech synthesis and most groups with normal vision was instructed and trained during special trials. The training of visual impaired subjects was made before testing during professional occupation or teaching, which has a component of speech synthesis. Duration of training varied from some months to some years from listener to listener. I t is possible, that training is connected with speech learning at all. Our experience indicated that this learning possible consists of two periods. The first period is rapid and short and another – slow and more prolonged. The first period continues about 10 min (if feedback is presented), and the second can lasts several weeks.

Before testing our group of listeners with normal vision we allowed to listen and analyze synthetic speech they heard about 10–20 min and therefore was made primary training at least. The second procedure in preparing listeners for testing is so called calibration. It means person's teaching for self-analysis or teaching to use self- considering as device or equipment for measuring.

All of listeners to be used in our testing had normal acoustic perception. All listeners were advanced in their native language (Lithuanian, Russian or Polish) and had no more or less obvious impairments of reading or writing. Listeners were monolingual (Russian), bilingual (Russian and Lithuanian), some of them trilingual (Polish, English, French or Germany in addition).

The number of subjects will be indicated in results.

Design. There were three periods in testing of synthetic speech intelligibility and acceptability: preparatory, intelligibility testing and acceptability testing. The order of these periods was standard and uniform for all listeners.

Testing was individual, the test performance was monitored and takes as long as 1.5–2 hours. All listeners were informed that their responses will be confidence.

During a preparatory period the listeners were trained, calibrated and after that they received complete instruction for testing. Thus the listeners were well familiar with goals of study and requirements for listening and evaluation. If it was needed, the preliminary studies there were training sessions for listeners. Real testing started only after experimental confidence that listeners all understood, learned and could give a stable responses. All responses had spoken form and were registered immediately after presentation of test material (letter, word or sentence).

The second period of testing is devoted to evaluation of intelligibility of synthetic speech. For Russian speech synthesis study were created 9 parallel test versions and for Lithuanian – 3 in such way as was described above. The first subtest (letters) in all versions had the same collection of letters but was in different pattern by randomization. The second (words) and the third (sentences) subtests in this parallel versions had different test material and different pattern after randomization.

Speech intelligibility test performance consists of different trials for each listener: Russian announcer, Russian synthesizer, Lithuanian announcer and Lithuanian synthesizer. Effects of local learning, fatigue, changes of attention and emotions, etc. were balanced by means of different versions of test and by randomization of trial order. In other words – before testing was created special experimental design for 50 listeners. This design told us what listener must be used, what test version must be presented, what speech (Russian or Lithuanian) in what form (synthetic or announcer) must be used, when test version must be presented (in first, second, third or fourth order) etc. So, this design help us to balance all experimental factors according requirements of psychological experimentation and allowed to test rapidly without delays and confusions.

Investigator recorded responses of a listener on a special paper or on a tape recorder. After testing the primary analysis was made and the main parameters of listener response were detected.

The further data processing was performed by special computer SPSS program and other software (statistical and functional analysis).

During the third period of testing listeners worked with speech acceptability and synthetic speech apply field *Questionnaire*. It was important that evaluation would be independent and without any influence from environment, other people or researcher. Listeners were asked to be careful not to hurry and well process own responses.

Test performance conducted in test room with normal environmental conditions where worked only two persons – researcher and listener. There were no any speech degradation versions or special device for speech signal processing to be used. Synthetic or natural speeches were presented at the level, which was the most acceptable for listener and it was determined during preparatory period. There were no any estimations of speech loudness in the test room. All versions of natural and synthetic speech were recorded on magnetic tape by means of the same equipment. Before each test magnetic tape recorder head was cleaned and quality of sound subjectively evaluated by researcher. Loudness during testing trials for synthetic and natural speech was equal. Tape recorder for speech reproduction was portable, which like most of visually handicapped. The male voices of Russian and Lithuanian professional announcers were recorded on magnetic tape at the audio studio in Moscow and Vilnius, where are producing master tapes for talking books for blinds.

The first experiment was designed to measure the intelligibility of currently used Russian synthetic speech and to compare it with corresponding characteristics of the Dolphin Company (Great Britain) Lithuanian version, RSP and LSP speech. A group of subjects representing Lithuanian population and including 20 blind and visually-impaired subjects and 28 visually normal subjects were investigated.

In the course of the second experiment, IRS intelligibility was measured and compared with corresponding characteristics of RSP and RS. In this case, also, Lithuanian population was investigated. It included 20 blind subjects and 20 visually normal subjects.

The purpose of the third experiment coincided with the purpose of the second one, only experiments took place in Moscow and involved native Russians. It was aimed at finding out whether the Lithuanian population subjects who know Russian evaluate intelligibility and acceptability of synthetic speech in the same way as native Russians who live in Russia.

2.3. *Indices and Data Processing*

Speech intelligibility was assessed by the rate of correct reproductions (CR) of the presented speech units (letters, words and sentences). Some speech units were reproduced only partially (with some changes, but in the same or close meaning). We named it partially correct reproductions. For reproductions of sentences we used additional index totally correct reproductions without meaning and word form distortions). Opposite index for first two is incorrectly reproduced or unrecognized speech units.

Speech acceptability as well intelligibility and applicability of synthesized speech was assessed by categorized answers to the questions or by 10-points scales of *Questionnaire*.

More than 50 speech intelligibility and acceptability indices were measured per each subject, and, after the primary data processing, statistical parameters (means, standard deviation, quotients of correlation, statistical reliability according to Student's test, etc.) were calculated by means of SPSS computer program.

2.4. Additional Research of Evaluation of Applicability of Speech Synthesis to Producing Talking Books

The purpose of this additional investigation was to find out how users are evaluating various talking books (traditional magnetic record in the tape recorder cassette, structured digitized talking book), various speakers (announcers), and application of speech synthesis to the producing of talking books. Two questionnaires were offered: the questionnaire aimed for measuring the quality of synthetic speech and use it for producing talking books and the questionnaire designed to measure characteristics of a structures digitized talking book.

The investigation was effected in two stages in Vilnius and Moscow. In Vilnius 40 non-expe-rienced young (17–22 years) subjects were interviewed. Their notion of a Russian talking book was rather vague as most of them had never come across either the talking book or the phenomenon of speech synthesis in general, and had no corresponding experience. Therefore the analysis will be founded on the data obtained in the course of investigating 12 subjects who were Muscovites employed with LOGOS enterprise and connected with talking book and synthetic speech technologies. Many of them were regular readers of such books, therefore they were chosen as experts.

The average age of subjects was 40 years. Half of the subjects had secondary education, and the rest of them – high technical or humanitarian education. Russian was the native language of all of them. The length of subjects' working with synthetic speech varied: 33 % had no previous experience in this field at all, 25 % only tried to, while 42 % used it often.

The investigation consisted of two separate stages, carried out either in uninterrupted succession, or with a break between the stages, the length of break being not less than one day (the break was necessary to have a rest, dinner or to arrange affairs connected with the official duties performed by subjects). In the first part, various literature: fiction, legal, technical in the form of extracts from various books (duration 15 min) presented in a male voice was offered to the subject's audition. All the texts involved in the questionnaire were presented in the following three forms: traditional talking book in a magnetic recorder cassette sounded by an announcer; magnetic record in a cassette sounded by means of currently used Russian speech synthesis; and magnetic record produced using an improved Russian synthesis.

In the course of the second investigation stage both questionnaires were offered to the subjects. The first questionnaire included 30 questions aimed at evaluation of the area of possible application of synthetic speech, 23 questions designed to reveal the traditional talking book user's evaluations, and 6 questions aimed at disclosing individual characteristics of subjects. The second questionnaire included 34 questions, their purpose was to evaluate the structured digitized talking book and personal characteristics of subjects.

The investigation was individual, lasted 3–4 hours and was carried out in a separate room protected from sound and light irritators. The rhythm of hearing was individual, the duration of breaks among hearings belonged to the subject's discretion. Besides the fragments of talkingbooks, voiced letter, word and sentence test aimed at objective evaluation of announcer's speech and of the two versions of Russian synthetic speech were presented. The answers were recorded by the investigator. The primary data processing was performed immediately after the experiment, in the evening of the same day. Necessary calculations were also made on the same day or a bit later.

2.5. Evaluation of Hungarian and Italian Synthetic Speech Acceptability

As was mentioned we had possibility to compare results of our investigation with the results of investigation of Hungarian and Italian colleagues, which used items 1–6, 10–17, 18, 19, 20, 21 and 27 from our *Questionaire/Questionnaire*. Hungarian subjects (12 visually impaired people) were interviewed by Andras Arato, Lashlo Buday and Teresa Vaspori. For evaluation were used two Hungarian synthesizers: *BraiLab* and improved version *BraiLab PC*.

In Italy synthetic speech acceptability was evaluated dr. Paolo Graziani in co-operation with the Italian Blind Union. For subjects (7 blind people, experienced in using of synthesized speech, excluding one) tape records of two synthesizers were presented. One tape record was made using synthesizer *Eloquens* developed by CSELT and another – using synthesizer *Audiologic TTS2*. *Eloquens* is mainly devoted to applications in telephone services of Telecom Italia. It presents a Windows application which is not yet used by blind people. *Audiologic* is one of the most diffused speech synthesizers among blind users. It is particularly appreciated for its quality. For this experiment the new version of this voice was used, improved both in quality and in flexibility for the correct interpretation of a text.

3. Results and Discussion

3.1. Intelligibility of Speech of Two Russian and Lithuanian Synthesizers

As can be seen in Figure 1A, according to CR (correct reproduction) parameter, RS (1st Russian synthesizer) speech intelligibility is significantly worse than in RSP (improved Russian synthesizer) speech intelligibility for letters (t = 30.31), for words (t = 23.0) and for sentencies (t = 22.10). In all cases $p > 0.005$ (Student test). As can be seen from CR parameter ratio (figures above histogram graphs), it is easy to be convinced that RS is 2.6 times as bad as RSP according to letter recognition. The same tendency is observed in case of words (1.4 times) and sentences (1.9 times).

It is evident that LS intelligibility is greater: for letters (t = 10.97), for words (t = 7.98) and for sentences (t = 5.33). The ratio being calculated, it is easy to see that, according to CR parameter, RS intelligibility is approximately 1.8 times worse than LS intelligibility for letters, 1.5 times for words and 1.4 times for sentences.

The same tendencies can be seen from the Figure 1B, where index of partially correct reproductions was used for assessment of speech intelligibility. These results allow conclude: The speech of both Russian and Lithuanian speaker, according to the number of correctly recognized stimuli, is more intelligible than Russian or Lithuanian synthesized speech. And this is no wonder, as the quality of both variants of synthesis is still clearly behind the natural speech.

3.2. Improved Russian Synthetic Speech Intelligibility in the Lithuanian Population

Figure 2 represents results of comparative evaluation of three kinds of Russian speech units (letters, words and sentences) tape recorded by

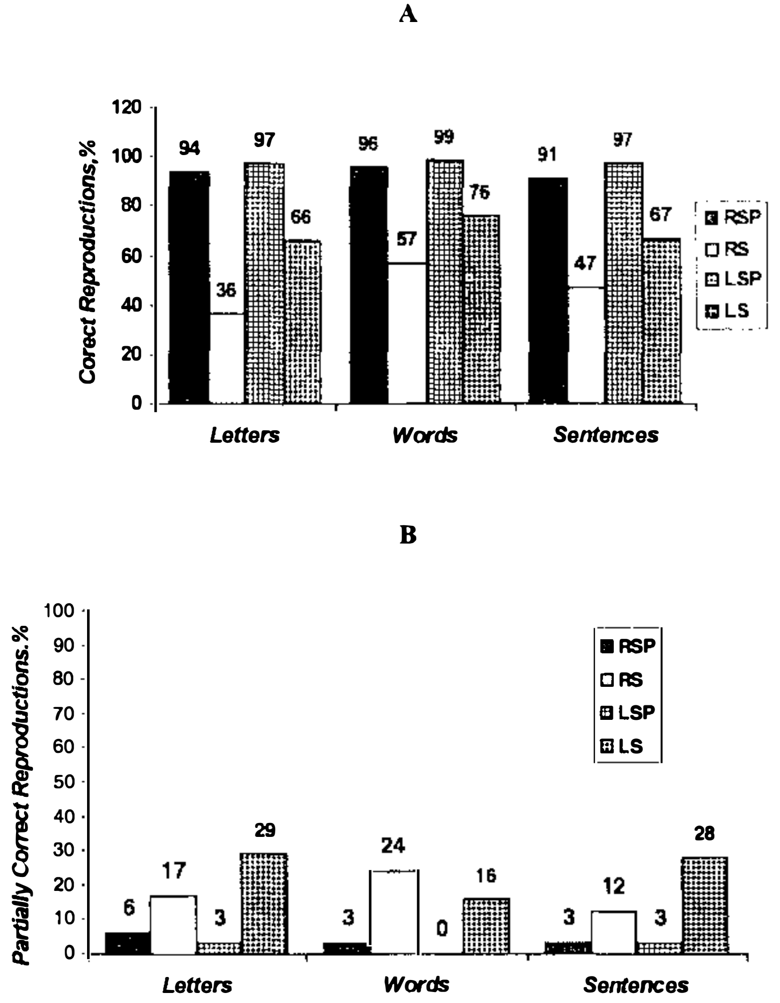

Figure 1. ***Number (in %) of correct (A) and partially correct (B) reproductions of speech units (letters, words and sentences) tape recorded by Russian announcer (RSP), first Russian synthesizer (RS), Lithuanian announcer (LSP) and Lithuanian synthesizer (LS)*** *(Data represent averages of Lithuanian subjects: 20 visually impaired and 28 visually normal subjects).*

the first and improved Russian speech synthesizers (RS and IRS) as well as by announcer (RSP). Evaluations were made by Lithuanian listeners (20 visually impaired subjects trained in perception of synthesized speech and 20 visually normal instructed subjects). All subjects were bilinguals, they knew Lithuanian and Russian. 20 visually normal subjects were selected among subjects which participated in the first experiment as they knew Russian better. In other words, in this way we tried to eliminate the small asymmetric deviation in favour of Lithuanian which we observed in the course of the first experiment.

In Figure 2 three indices of efficiency of perception tape recorded speech units are presented: correct reproductions (A chart), partially correct reproductions (B chart) and totally correct reproductions (C chart). According to all three indices intelligibility of RSP speech units is higher than IRS and RS. The quality of the improved Russian synthesis is better than the quality of first Russian synthesis. Besides, this improvement is achieved rather by improving physical characteristics of acoustic equivalents of letters than by making the synthetic speech more available to human verbal mechanisms, though even here a slight improvement can be observed. This makes performance of separate letter analysis worth-while in order to find out what particular sounds contribute to the general improvement of synthetic speech intelligibility most.

Seeking to prove once more the correctness of our conclusion about the increase in the quality of improved synthesis being essentially achieved at the expense of the improvement of sound quality in sentence sub-test we distinguished another dependent variable – index of totally correct reproductions (correct reproduction of meaning and correct reproduction of words consisting sentence). Figure 2 C chart shows that total correctness of reproduction is a little less (90 %) in comparing with correct reproductions of meaning (97 %, Figure 2 Chart A). For RS and IRS indices of totally correct reproductions is 57.78 %, and 62.22 %. It follows that the new synthesis, according to this parameter surpasses the old one only by 4.44 %.

In conclusion, we would like to remark that here also all standard calculations were made. Tendencies provided here are statistically reliable.

3.3. Intelligibility of Improved Russian Synthesis in the Russian Population

As has been mentioned, in order to verify once more whether the above-mentioned tendencies are really applicable to native Russian residing in Russia, we made an additional experiment in Moscow using the same investigation methods. Subjects were employed with the LOGOS institution, so they were professionally-related to talking book technology. Some of them were highly experienced in the use of and work with RS synthesis variant. As can be seen in Figure 3, according to the quantity of correctly reproduced speech units the new IRS synthesis is better both for letter and word parameters. If RSP speech is taken as the point of departure, then for letters, RS intelligibility equals to 57.28 %, IRS intelligibility being 76.04 %. It is obvious that here IRS intelligibility increases by 18.75 %. It is a little less than the analogous index in the Lithuanian population (23.71 %, see Figure 2).

Corresponding evaluations of synthetic speech intelligibility for words are as follows: RS – 76.77 %, and IRS – 81.82 %. Intelligibility of the new synthesis is improved by 5.05 %. In the Lithuanian population this value equals to 11.22 %. Thus we can observe the same tendency here too, but here it is less expressed.

The most interesting results were obtained in the sentence reproduction test (Chart A on Figure 3). It is obvious that here also intelligibility of RS synthesis is the highest, while the natural speech and IRS even lag behind it a bit (1 %). From first sight, the results seem to be paradoxical: the improved synthesis cannot be better than the natural speech, can it? The Chart C (Figure 3) shows that totally correctly reproduced sentences best for the announcer's speech, then follows the first Russian synthesis, and then the improved Russian synthesis. So,

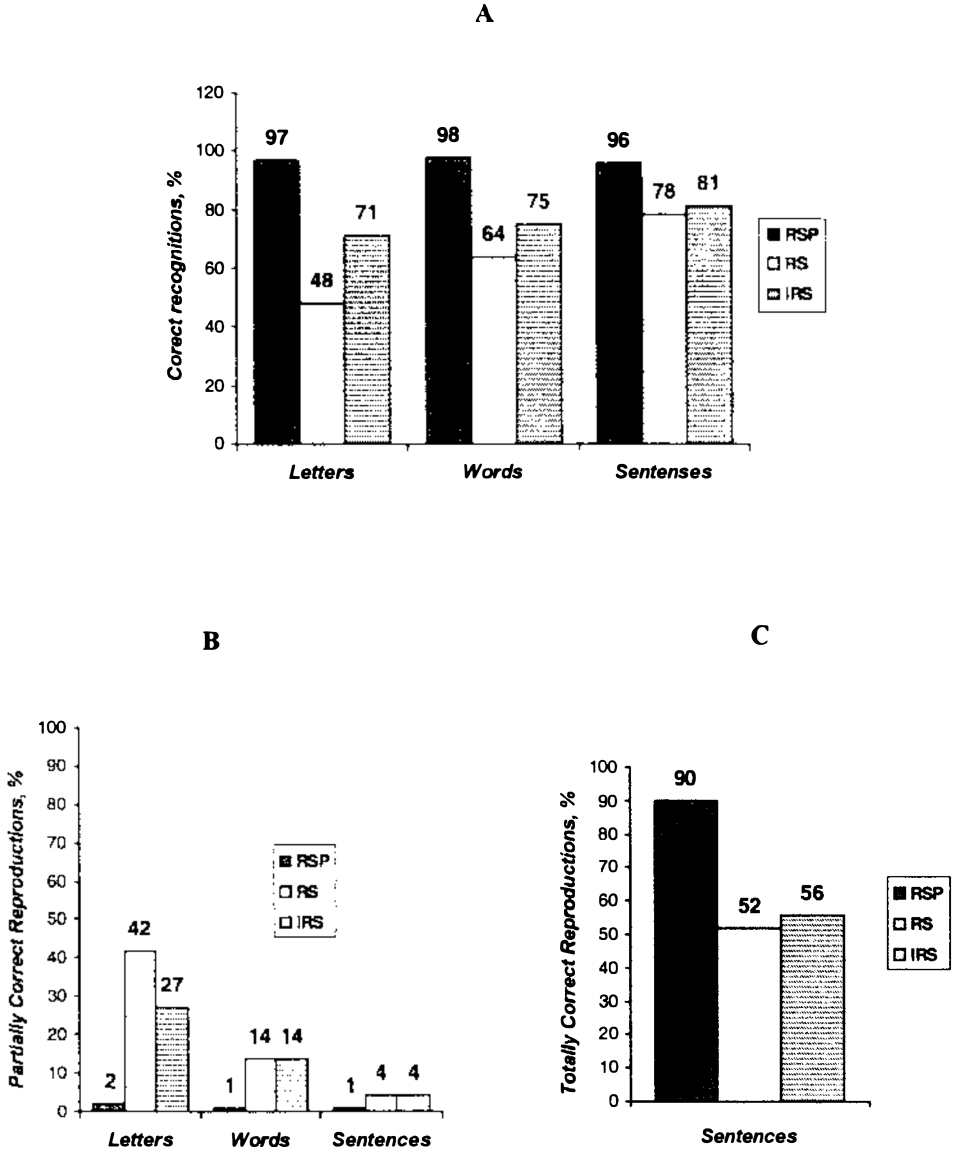

Figure 2. ***Number (in %) of correct*** *(A),* ***partially correct*** *(B)* ***and totally correct*** *(C* ***only for sentences)*** ***reproductions of letters, words and sentences tape recorded by Russian announcer*** *(SP),****first Russian synthesis version*** *(RS)* ***and improved Russian synthesis*** *(IRS) (Group of 20 visually impaired and 20 visually normal subjects of Lithuanian population)*

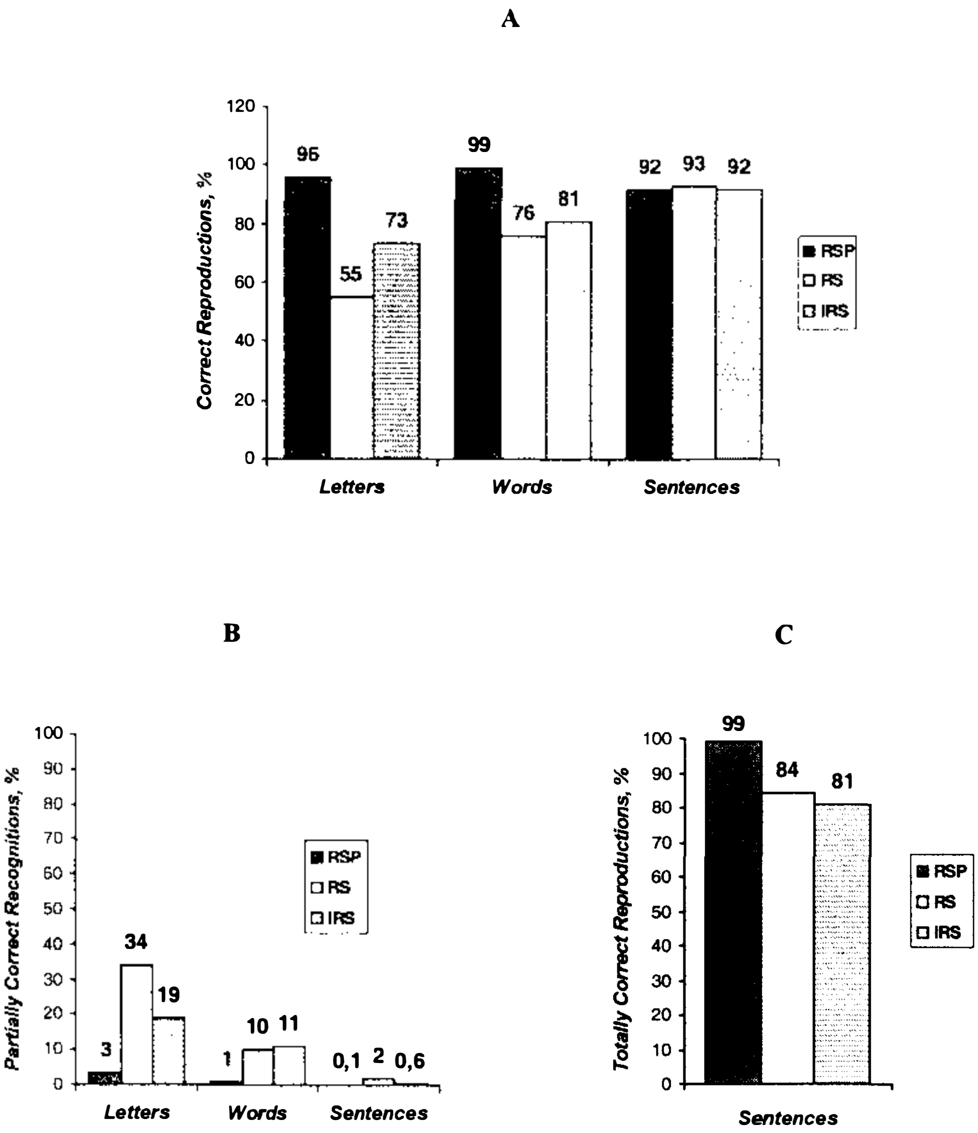

Figure 3. ***Number (in %) of correct (A), partially correct (B) and totally correct (C only for sentences) reproductions of letters, words and sentences tape recorded by human speaker (RSP), first Russian synthesis version (RS) and improved Russian synthesis (IRS)*** *(Group of 9 visually impaired and 3 visually normal subjects of Russian population)*

thc ncw synthesis is worse than thc new one according to this parameter, too. How can this paradox bc explained?

A more detailed and attentive analysis revealed that this paradoxical tendency had been brought about by the subjects who were exceptionally experienced in operating RS version. The quantity of correctly reproduced speech units in their case clearly depended on the level of stimuli meaning. The greater meaning of a stimulus, the better RS intelligibility for it, even better than the intelligibility of the announcer's speech. Thus, we encounter here the phenomenon called "training". It could be defined more precisely as a specific adjustment of verbal mechanisms to a corresponding synthesis version which appears as a result of long experience.

This phenomenon indicates distinctly that an objective evaluation of a synthesis version needs not only professional evaluators with a long period of experience in working with synthesis. Their evaluations can differ from those given by non-experienced potential workers without a corresponding experience. Consequently, both types of evaluators are needed to achieve objective evaluation. Therefore, in the analysis we used mixed groups of subjects. This phenomenon should be borne in mind not only in the evaluation of new versions of synthesis, but also in organizing of training of subjects non-experienced in speech synthesis.

3.4.Improved Russian Synthesis: Characteristics of Sound

As has been already mentioned, the general analysis of results made us draw the conclusion that new synthesis IRS improvement mostly depended on sound synthesis, the improvement in comprehension of verbal information having contributed to it in a lesser degree. So now we shall try to see how the general sound synthesis was improved in the new IRS version and what particular sounds were involved.

Let's start with the analysis of recognition of sounds uttered by a human speaker. The Lithuanian population used in the second experiment can be taken as an example. In total 1120 separate sounds were presented to all subjects. The primary data analysis showed that only 27 sounds failed to be recognized correctly. In other words, as many as 97 % of all separate sounds (phonemes of letters) were recognized correctly. "TS" sound resisted recognition most (5 times it was comprehended as "S", "F" ,5 – "S") and "ZH" (4 – "Z"). "I", "SH", "A", "R", "YE", "V", "Z", "D", "YA", "K", "L", "YU", "CH". "KH", "U" and "H" sounds were among best recognized ones.

RS versions of the currently used Russian version synthesis in the said group of subjects had only two sounds, namely "O" and "I", that were recognized 100 %. All other sounds had more often heterogeneous and less often homogeneous incorrect recognition. "F" sound can be mentioned as an instance of homogenous incorrect recognition, it was taken for "S" sound as many as 24 times. Homogeneous incorrect recognition was most strong and most often. "V" sound was the most difficult to recognize (37 mistakes, it was usually taken for "R", "T", "P", "F", etc.). "KH" sound (31 mistakes, usually taken for "K", "G", "F"). "YU" sound (29 mistakes, usually taken for "I", "U"). On the whole, though there were mistakes in the recognition of all sounds, the above-mentioned sounds were misrecognised most often.

In the improved synthesis version IRS only 4 sounds had many heterogeneous mistakes. Those were "P" sound (28 mistakes), "V" sound (25 mistakes) and "KH" sound (23 mistakes). Other sounds usually scored 3–10 mistakes. On the whole, here, with the exception of "KH" sound, homogeneous mistakes became less in number at the expense of heterogeneous mistakes. "YE" sound was synthesized best. "YA", "CH", "O", "U" "S" and "YU" sounds scored only several mistakes each. Thus, two opposite tendencies can be observed here: according to the general quantity of mistake reduction, this version is tending towards the natural speech, but according to the homogeneity of mistakes, it moves away from the natural speech. As the first tendency is clearly dominant, the general resultant in the new version shows a tend to improve. In future, the improvement of IRS synthesis variant should develop in two directions. The first has been already tried, it gives general reduction of mistakes. The second way of reduction of heterogeneous mistakes has not been tried; judging from achieved results, it seems to lead in the opposite direction away from the natural speech.

Summarizing the analysis of the quality of verbal sound synthesis in the Russian synthesis, we would like to note that both versions share a common shortcoming noticed by many subjects and characterized by some of them as a very disturbing hindrance to the acceptability of versions. The said shortcoming represents the absence of "YO" sound in both synthesis versions.

3.5. *Intelligibility and Acceptability Correlation of Improved Russian Speech Synthesis*

The next step in our analysis is the investigation of correlation between various parameters. Here, also, we'll make use of data obtained in the course of the second experiment. The subjective (scaled) intelligibility of announcer's speech was very poorly correlated with the correct recognition of letters (r = –0.13), but better with the same parameter in words (r = –0.42), sentences (r = –0.34) and totally correct reproductions of sentences (r = –0.22). This enables us to presume that separate speech sounds make little impact on the intelligibility of natural speech, while the meaning of words and sentences is the most powerful factor in this respect. In the first Russian synthesis version RS the significance of sounds rises a little (r = 0.15), the same can be said especially about the correct recognition of words in a sentence (r = 0.52) and the significance of totally correct reproductions (r = 0.46). This means that the articulation of synthesis sounds is a really important factor, however the sentence is undoubtedly most important thing that helps bring about the result. In other words, chances of correct word recognition are increased if they are presented in a sentence, therefore it is only natural that the impact of the general idea of a sentence manifests itself very powerfully wherever general speech comprehension is concerned.

In the improved speech synthesis version all above-mentioned correlation coefficients are almost the same as in RS, except one: the subjective intelligibility of IRS is again highly correlated with the correct word recognition (r = 0.49) which is characteristic of the natural speech. It may be possible that subjects start to regard the improved synthesis as a natural speech, but the preserved high coefficient of correlation with the correct sound recognition (r = 0.18) reminds us that it is just a speech synthesis, and nothing more.

Correlation of subjective speech acceptability with various parameters of its objective intelligibility are interesting too. In the natural speech of announcer all acceptability correlation with the correct sound recognition (r = 0.14), correct recognition of separate words (r = –0.04) and correct recognition of sentences (r = -0.02) and totally correct recognitions of sentences (r = 0.000) are very small.

In the improved version Russian synthesis, correlation between subjective acceptability and objective intelligibility are almost the same as in the natural speech, except one: the correlation between acceptability and correct recognition of sounds rises.

It would be possible to think that those correlation also indicate the progress of the new synthesis towards the natural speech, however, the rise of the said correlation coefficient makes us be cautious.

The age and gender of subjects showed statistically unreliable and very low correlation with the objective comprehension of all speech variants, the comprehension being measured according to the number of correctly recognized stimuli. This makes us think that the age and gender of subjects has little influence upon objective parameters of synthesis.

3.6. Characteristics of Russian Synthetic Speech Application

The first three questions in the questionnaire were designed to evaluate the comprehensibility of the natural and the synthetic speech by means of a 10-point scale. The data indicate that the announcer's speech was most intelligible – 9.37 points. The currently used Russian synthesis rated average 5.58 points, while the improved synthesis – only 4.96 points. Though the difference in evaluation of comprehensibility of both syntheses is not statistically significant, the lower evaluation rate received by the improved version of synthesis is a bit unexpected. A more detailed analysis shows that an exceptionally low evaluation of the improved synthesis is given by the subjects with extensive experience with the previous version of synthesis. Thus, the first cause of such evaluation represents the training phenomenon described in literature, the said phenomenon lies in verbal speech comprehension mechanisms getting adapted to corresponding speech.

The second cause of such evaluation emerges from the analysis of acceptability data and can be called a criterion shift. Here, also, the highest rating was given to human voice – 8.96 points. The first synthesis rates 5.27 points, and the improved one – 5.71 points. Thus, the improved synthesis is more acceptable to subjects. In other words, they view it as more natural than technical, consequently, their judgements are more strict. Or, to put it more precisely, the old synthesis is viewed by subjects as a rather decent instance of a robot's speech, while the new synthesis – as a poor variant of human speech. This criterion shift is an obvious criterion of progress in the improved version of Russian synthesis.

All subjects also indicate various shortcomings of both the natural and the synthetic speech. The natural speech of the announcer wanted more distinct pronunciation of separate sounds, especially in the beginning of a word or a sentence. Some subjects were annoyed by expressed pronunciation of high-frequency components in fricatives and affricates.

Subjects found even more faults with the synthetic language. They demanded more "human" sounding, naturalness and more distinct pronunciation in the currently used Russian synthesis. The "metallic" or "tin" sound quality of the voice was emphasized. Absence of human intonation and prosody was stressed.

As for the advanced Russian synthesis version, subjects' demands bore a more verbal character. Presence of accent was often indicated: a characteristic Muscovite pronunciation ("a" instead of "o" in certain positions) on the one hand, and a Polish accent with the characteristic "psh"-type sounds. Absence of "yo", hasty articulation of certain sounds, outside non-speech sounds and big efforts needed to comprehend speech were stressed. There were wishes relating to overtones (low-frequency, better discrimination of sounds) was demanded.

Striking differences between the natural and the synthetic speech manifest themselves in the subject's wishes relating to preferable type of literature suggested for hearing. It was mainly fiction (77 %) and a bit of technical literature and literature of other types (7.7 % respectively) that was hearted most willingly if reproduced in the natural voice of announcer. However, the use of synthetic speech in talking book production received a quite different evaluation. According to subjects, synthetic speech was absolutely nonapplicable in fiction, both prose and poetry (0 % respectively), being most applicable in technical texts (30 %). Further, scientific (25 %), political (20 %), popular literature and journalistic writings (10 % respectively) and other literature (5 %) was mentioned.

Improvement of intelligibility of synthetic speech in the course of hearing is an interesting phenomenon. Besides, the poorer is the subject's experience in working with synthesis, the more striking is the said dynamics. Even 33 % of the subjects consider this improvement to be rapid, 25 % of them consider it to be of medium speed. One third of all subjects do not notice any dynamics (33 %), some 9 % said that in the course of listening to synthetic speech its intelligibility decreases. The differences in dynamics may be useful in the organization of synthetic speech training courses. The subjects have gained such experience mostly by working with the currently used Russian synthesis (50 %); only a part of them (25 %) acquired the experience when working with other syntheses. Some 25 % of subjects did not have any serious experience in working with synthesis.

The duration of synthetic speech usage also varied in subjects. The average duration was 2.57 years: 8.3 % of subjects dealt with synthesis 6 years, 8.3 % – 5 years, 33 % – 4 years, 17 % – some 2 years, and 33 % of subjects did not have any substantial experience in synthesis.

The frequency of usage of synthesis also differed. It was principally related with the subject's profession. It is not surprising, as voice synthesis technology has just been introduced into practice. Approximately one third of subjects used synthesis every day or at least once a week (33 %). Some 25 % of subjects worked with synthesis once a month, 9 % of them had come in touch with synthesis only once, and 33 % of subjects came across it quite by chance.

It is necessary to notice, that quite recently synthetic speech has been employed in hearing of various texts. As many as 50 % of subjects, by means of synthesis, read technical texts and other information, 18.3 % – read juridical texts, while some 41.7 % of subjects do not use synthesis in hearing various texts at all.

Subjects had different opinions about the area of application of synthetic speech. Some 14 % of subjects did not know how synthetic speech could facilitate their life. Some 25 % thought that synthesis would be used in reading books, 15 % – in professional activities; 10 % – in household activities and 5 % – in studies. 15 % indicated that synthetic speech could be probably applied in the information area, and 5 % of the subjects were real synthetic speech enthusiasts. They thought that synthetic speech could be applied wherever the text was present. Part of blind subjects (10 % of all population) forecast its application in the area of technical service (e. g. telephone).

All subjects noticed a number of general shortcomings both in the current and in the improved variant of synthesis. Many subjects did not like the absence of "yo" sound in synthesis, the "yo" being replaced by "e". Bad pronunciation, outside sound effects, artificiality, indistinct articulation, stability of intonation were mentioned as comprehension-disturbing factors. The improved variant of synthesis, as has been mentioned, was viewed as more "human".

Stress is a major drawback of synthesis. Though 33 % of subjects did not complain about stress mistakes, 55 % were annoyed by them when listening to certain texts, and 17 % said that stress mistakes interfered with hearing of all texts.

Almost all subjects (91.7 %) wished the speed of synthetic speech could be controlled, and only 8.3 % said the natural tempo in synthesis would satisfy them. A rather distinct tendency was revealed in the analysis of answers to questions about preferable voice of synthesis. Man's voice was most preferable (50 % of all subjects). Woman's voice was considered to be more emotionally pretentious, therefore it was less preferable in the respect of text intonation. Some 25 % of subjects did not have any clearly expressed preferences. The same proportion of subjects said that the choice of voice depended on the character of text. Not a single subject chose a child's voice.

Speaking of blind people, the most acceptable way of reading, in the opinion of our subjects, would be the reading of talking books (30 %). Reading in Braille was suggested by almost the same number of subjects (25 %). Only 20 % of subjects indicated that reading by means of synthetic speech would be most suitable to a blind person. Such a rather low evaluation of synthesis may be caused by the fact that the advantages of reading (reproducing) of a structured digitized talking book were unknown to the majority of subjects. It is possible that many subjects imagine that synthesis is nothing more than an artificial voice's substitution for a natural one in a magnetic recorder's cassette. In the same way (i. e.20 %) the significance of speaking mass media (radio, television) to a blind person was evaluated. Only 5 % do not reject the possibility of offering the blind some other ways of reading.

42 % of subjects who often use synthesis considered themselves, subjectively, to be synthetic speech users. Some 25 % of subjects tried synthesis just to satisfy their curiosity, 33 % of them did not come in touch with it. Thus we can see that more than a half of LOGOS employees have something to do with synthesis. This substantiates the subject's belief that in future the frequency and intensity of the use of synthesis will grow. This fact indicates that speech synthesis is regarded as a progressive and innovative thing.

3.7. Users' Views on a Traditional Talking Book: Russian Population

Traditional talking books have firmly held the leading position in the reading habits of blind and visually-impaired people until now. It is important to know how the traditional talking book (tape-recorded) is viewed by its user and its producer.

The analysis of results offered by the first part of the questionnaire show that some 67 % of subjects (visually normal, blind and visually-impaired) are regular readers of traditional talking books. Some 25 % of subjects have hearted more than 10 books, and 8 % of them confessed that they had no serious experience in this field. As has been mentioned, this way of reading is the most popular. Therefore we shall engage now in a more detailed analysis of usage of the traditional talking book.

First, quite unlike the synthetic speech, the traditional talking book is mostly used in the reading of fiction (prose – 77 % of subjects). Scientific talking books are scarce, even 35.3 % of subjects complained about it. Some 17.6 % of subjects would like a greater choice of fiction (prose). Further goes technical literature, other literature, and answers without respondent's separate opinion (11.8 % respectively). It is important to emphasise that all subjects do not think the traditional talking book to be an adequate means of reading poetry. When speaking, subjects stressed that the reading of poetry required individual alternative speed and purely personal emotions. Poetry read by an announcer or an actor was often rejected because of emotions expressed by them. In this matter subjects are concerned with their own emotions. Subjects say that good reciters/nonexistent word/(announcers or actors) are extremely rare.

When asked a question about the total number of issued talking books, as many as 50 % of subjects answered that the numberwas undoubtedly insufficient. Besides, 25 % said that the number was insufficient and only one fourth (25 %) of them thought that it was sufficient.

Almost all subjects indicated certain shortcomings in the current technology of traditional talking books. The greatest complaints were directed towards the quality of currently used magnetic recorders (50 % of respondents). Another 25 % of respondents complained about the control of "reading" of current traditional books. With the introduction of computer-based reading and the availability of good speech synthesis, the problem of bad equipment would be settled radically. Some 80 % of subjects complained about the speaker's voice and speech, similar proportion of subjects could not say anything definite about it.

Future prospects of traditional talking books were rather clear to subjects. Some 75 % of them said that blind people would use them long time, and only 25 % thought that in the nearest future they would be replaced by computer-based reading methods. Such great evaluation of prospects of traditional talking books was determined by economic factors. The greater part of subjects thinks that an ordinary blind Russian cannot afford buying a personal computer yet. However a possibility to organize the reading in some other way is not rejected (computer rent, special reading-rooms, etc.).

Many subjects indicate various, and important, in their opinion, shortcomings of the current traditional talking book. The shortcomings do not have a clear statistical general tendency, but they are rather important individually. Lack of lucidity in the criteria according to which literature is chosen in talking books production was mentioned. Certain listeners are annoyed by over exaggerated emotionality of the speaker, too great distance separating the book from its reader and too long queues of readers seeking a popular book. Greater diversity in genre (e. g. more children's literature, teaching aids and scientific literature) was demanded. Rather often complaints about the quality of magnetic tape, cassette and recording or reproducing equipment were heard.

3.8. Evaluation of Structured Books: Russian Population

Digital component being introduced into the talking book technology, the production of structured digitized talking book would be the next stage in its development. To accomplish it, besides a computer and synthetic speech, a reader CD and a special recording control and text compression program are needed. Therefore, it is important to know what people know about them, and how they appraise them. A special questionnaire aimed at revealing characteristics of structured digitized talking book was used. The investigation was created in Moscow. The analysis will be founded on the data obtained in the course of investigating 12 subjects who were Muscovites employed with LOGOS enterprise and connected with talking book and synthetic speech technologies. Many of them were regular readers of such books, therefore they were chosen as experts.

Only 41.7 % of respondents had formed a true notion of the concept of "structured digitized talking book". The rest 58.3 % had a vague idea about it. The same proportion emerged from the answers to the questions about the main parts of digitized book, i. e. its chapters and paragraphs. Footnotes and picture descriptions (25 % positive answers) and hyperlinks (16.7 %) were even less clear to the respondents.

Analysis of answers to the questions about basic characteristics of digitized book produces similar results. The graph describing the book's structure was understood by 25 % of respondents; graph mode identifier – by 16.7 %; reading unit – by 16.7 %; book marker or commentary notes – by 33.3 %; "sighted" page number – by 33.3 %; possible reading modes – by 25 %; automatic book markers – by 33.3 %; navigation and current node – 33.3 %; sound effects connected with speech synthesis – by 16.7 %; type prediction – by 16.7 %. Thus, it is possible to state that the structured digitized talking book is known to less than 1/3 of respondents, representing, on the whole, the participants in the creation of this new technology. Most of subjects had a vague idea about the concept of structured digitized talking book, though they heard of it, it attracted them, and their positive attitude towards it could be felt. Such understanding has good chances of developing in all directions with the further evolution of digitized recording studio.

About 33.3 % of respondents say they can start the reading program themselves; 8.3 % seek somebody's help in doing it; and 58.3 % of subjects are sure they cannot do it themselves yet. Similar answer is given also to the question whether the subject can read the digitized book himself/herself: 33.3 % can read it themselves; 8.3 % can read it onlywith the help; 58.3 % think they'll never manage it. The obtained data speak in favour of special training to be offered to prospect users of digitized book reading program. Some 25 % of subjects think that the existing reading program is not ideal, that it should be solved in some other way. While the rest 75 % cannot say anything definite about the adequacy of the program. This fact also speaks for the necessity of training.

Answering the question: "What is very good in the digitized book?", subjects enumerate a number of its advantages. The convenience of findingone's bearings in the text is mentioned, the compact information recording method, the long duration of high-quality information storage, the possibility to get information from any part of the book is emphasized. Also, the possibility to structurize and to get one's bearings in the text quickly, and the speedy search are mentioned. These characteristics are of strategic importance, they determine future distributionof this technology.

Some of the subjects indicate also certain undesirable characteristics of the structured digitized talking book. Users are taken aback by a possibly high price of the book. They think that many redundant functions have been included in the menu of command program. Subjects demand various synthesis programs designed for various languages that could be run on one computer. Subjects are concerned with user training difficulties, book library and storage peculiarities, as well as with the reader's chances to get information quickly. A possibility to enter internet by means of a magnified font in the course of reading is desired. A wish to simplify the reading program is obvious, the present intricacy of the program is repulsive to its prospect users.

Almost a half of subjects (41.7 %) view their chances of mastering the reading program optimistically; they think they can perform any service after several hours of training. Precisely the same proportion of subjects is more reserved: they think they will be able to perform only the basic service after several hours of training. Only 16.6 % of subjects think that more time and more efforts are needed to master the reading program. Almost a halfof subjects (41.7 %) say that they take delight in using the book. Only 8.3 % disliked the book, and 50 % were undecided. These figures also indicate the positive attitude of the user towards the digitized book.

3.9. Comparison of Hungarian, Italian and Russian Synthetic Speech Quality

Answers Italian, Hungarian, Lithuanian and Russian subjects to some questions (numbers 1–6, 10–17,18, 19, 20, 21 and 27) of main *Questionnaire* are presented (tables 1, 2, 3, 4, 5, 6). We used only questions which answers can be expressed in quantitative form.

First of all we offer to consider a subjective listener's opinion on the intelligibility and acceptability of the natural and synthetic speech.

Natural speech intelligibility scores obtained near highest evaluation (8–10 points), most intelligible is Hungarian and less – Italian announcers (1, 2, 3 items of *Questionnaire*). All versions and all languages speech synthesis accepted considerable lower intelligibility scoring (1–8 points) with increased dispersion. The improved versions were estimated as more subjective intelligible for Hungarian and Lithuanian synthesizers (improvement 1.7 and 2.1 respectively) and less intelligible – for Russian 2nd synthesizer (an objective intelligibility scoring, as we mentioned above, in this version obtained strictly opposite direction). It may be reasonable to assume that Italian listeners used enlarged criterion for natural speech intelligibility and Russian listeners were attached by synthetic speech of lower quality.

Acceptability studies showed more homogeneous scores at all. Natural speech more enjoyed Hungarian listeners and Italian were more critical listeners again. All versions of synthetic speech were judged as less acceptable than natural but after improvement most of listeners changed their mind and 2nd synthesizer was accepted as more acceptable then 1st one. In general all

*Table 1. **Averaged subjective judgement (10-point scales) of natural speech and synthesized speech** (In parantheses the minimal and maximal values are given)*

Mode of speech	*Hungarian*	*Italian*	*Russian*	
			Moscow	*Vilnius*
	(n = 12)	*(n = 7)*	*(n = 12)*	*(n = 40)*
10-point scale "***very bad – very good***":				
1.Natural speech	9.7 (9–10)	8.6 (8–10)	9.4 (8–10)	9.4 (8–10)
2. 1st synthesizer	6.8 (5–8)	6.7 (6–7)	5.6 (3–8)	4.5 (2–7)
3.2nd synthesizer	8.5 (7–10)	6.7 (5–9)	5.0 (3–7)	6.6 (3–9)
10-point scale "***very unacceptable – very acceptable***"				
4.Natural speech	9.9 (9–10)	8.7 (7–10)	9.0 (6–10)	9.1 (5–10)
5. 1st synthesizer	6.3 (3–8)	6.3 (5–8)	5.2 (3–9)	3.8 (1–8)
6.2 nd synthesizer	8.4 (5–10)	6.7 (6–8)	5.7 (4–7)	6.5 (2–9)

The 1st and 2nd synthesizers means: Hungarian study – one of former *BraiLab* synthesizers and new one *BraiLab PC*; Italian study – *Eloquens* and *Audiologic*; Our study – Kovax (1st Russian) and improved of Russian synthesizer. Results of evaluation of the Lithuanian synthesizer are not included in the Table

Table 2. ***Distribution of answers (in per cent) of respondents to the question "What texts in your opinion could be reproduced by synthesized speech?"***

Mode of speech	*Hungarian*	*Italian*	*Russian*	
			Moscow	*Vilnius*
	(n = 12)	(n = 7)	(n = 12)	(n = 40)
Fiction	8.3	42.8	0.0	27.5
Poetry	0.0	0.0	0.0	5.0
Technical information	91.7	85.7	50.0	72.5
Scientific literature	75.0	57.1	41.7	55.0
Political texts	66.7	57.1	33.3	32.5
Publicistic writings	33.3	57.1	16.6	27.5
Popular articles	91.7	71.4	16.6	27.5
Other	100.0	0.0	8.4	2.5

Table 3. ***Distribution of answers (in per cent) of respondents to the question "How quickly did Your understanding of the synthesized speech improved?"***

Mode of speech	*Hungarian*	*Italian*	*Russian*	
			Moscow	*Vilnius*
	(n = 12)	*(n = 7)*	*(n = 12)*	*(n = 40)*
Very quickly	50.0	0.0	0.0	0.0
Quickly	16.6	71.4	33.3	20.0
At medium speed	33.3	28.6	25.0	62.5
Did not improve	0.0	0.0	33.3	12.5
More listening, less understanding	0.0	0.0	9.0	2.5
No appraisal	0.0	0.0	0.0	2.5

Table 4. ***Distribution ofanswers (in per cent)of respondents to the question "What speech synthesizers have used?"***

Mode of speech	*Hungarian*	*Italian*	*Russian*	
			Moscow	*Vilnius*
	(n = 12)	*(n = 7)*	*(n = 12)*	*(n = 40)*
BraiLab Basic	83.3			
BraiLab Plus	50.0			
BraiLab PC	100.0			
PC Robot	8.3			
Dectalk	25.0			
Difon 2		57.1		
Audiologic		42.8		
Kovax			50.0	
Other Russian synthesizer			25.0	
Lithuanian (Dolphin) (mostly only trial)				40.0
English synthesizer				12.5
Never used or tried		28.5	25.0	47.5

Table 5. ***Distribution of answers (in per cent) to the question "How long have You been using the speech synthesizer?"***

Mode of speech	*Hungarian*	*Italian*	*Russian*	
			Moscow	*Vilnius*
	(n = 12)	*(n = 7)*	*(n = 12)*	*(n = 40)*
Haven't experience			33.3	74.3
Several hours				7.7
One month				10.2
One year	8.3			7.7
2–5 years	8.3	100.0	16.6	
5–9 years	75.0		50.0	
10 and more years	8.3			

Table 6. ***Distribution of answers (in per cent) of respondents to the question "How often do You use the speech synthesizer?"***

Mode of speech	*Hungarian*	*Italian*	*Russian*	
			Moscow	*Vilnius*
	(n = 12)	*(n = 7)*	*(n = 12)*	*(n = 40)*
Daily	66.6	57.1	16.6	7.7
Once in several days	8.3	0.0	16.6	0.0
Once a week	8.3	14.2	25.0	10.2
Once a month	16.6	0.0		

listeners detected an increased acceptability of improved synthetic speech versions.

Eight items in "*Speech Synthesizer Appraisal Form*" were devoted to detect and to specify the fields of applications of speech synthesis (Table 2). It is obvious that technical information and scientific literature were most relevant for Hungarian, Italian and Russian synthetic speech. Poetry and fiction obtained an opposite evaluation – they were less to be wanted for speech synthesis. So, we can conclude that logistic information without emotional component is more acceptable in synthetic speech application.

Distribution of answers of respondents to the question about changes in the synthetic speech understanding during listening trial showed considerable learning and training (table 3) processes. It is obvious, that these processes improved subjective intelligibility of synthetic speech for Italian and Russian listeners "at medium speed" or "quickly". A half of Hungarian listeners (50.0 %) reported this improvement performance to be changed as "very quickly". Such high percent of rapid changes in this population on one hand can be determined by enriched speech synthesis experience (Hungarian listeners participated in the speech evaluation more often than listeners from other countries) or by criterion shift. On another hand, enlarged percent of "did not improve" or "more listening, less understanding" answers in Russian and Lithuanian population was high related with they relatively small experience in synthetic speech quality evaluation.

Most of synthetic speech consumers used two versions of synthetic speech in their professional or daily activity (Table 4). So, Italians like *Difon-2*, *Audiologi*, Russians – *Kovax* and other Russian version, Lithuanians – *Dolphin* (adapted) and some of English synthesizers, only Hungarians chosen 5 versions, but as professional evaluators they work with *BraiLab*. We can assume, that in real life synthetic speech consumers (most of them are blind) tended to use only one or two synthetic speech versions.

Consuming time of various speech synthesis versions demonstrated considerable dispersion (Table 5). Most prolonged using time reported Hungarian and Italian listeners. Lithuanian and Russian consumers groups were small and not widespread, so they only started synthetic speech implementation at present moment. Partially it reflects how much blinds and other consumers use synthetic speech in daily life, professional activity and teaching.

This conclusion is confirmed by answers of respondents to the question "How often do you use the speech synthesizer?" (Table 6). The most enthusiastic daily speech synthesis users were detected Hungarian and Italian listeners, but Lithuanian consumers were youngest one.

4. Conclusions

1. Both Russian and Lithuanian speech units generated by announcer is more intelligible than Russian or Lithuanian synthesis. Quality of both variants of synthesis is still clearly behind the natural speech. First version of Russian synthesis is far worse, worse than Dolphin Company Lithuanian version.

2. Intelligibility of speech units generated by improved Russian synthesizer is higher than speech units produced by first Russian synthesizer. But both of synthesizers are far behind the natural speech. These tendencies were confirmed with Lithuanian and Russian subjects. Corresponding evaluations of synthetic speech intelligibility for words are as follows: RS – 76.77 %, and IRS – 81.82 %. Intelligibility of the new synthesis is improved by 5.05 %. In the Lithuanian population this value equals to 11.22 %.

3. Study of characteristics of sound synthesis shows that two opposite tendencies can be observed: according to the general quantity of mistake reduction this version is tending towards the natural speech, but according to the homogeneity of mistakes, it moves away from the natural speech. As the first tendency is clearly dominant, the general resultant in the newversion shows a tendency to improve.

4. Correlation between intelligibility and acceptability of speech deals possibility of thinks those correlation also indicate the progress of the new synthesis towards the natural speech, however, some correlation coefficient makes us be cautious.

5. The data obtained by *Questionnaire* indicate that the announcer's speech was most intelligible – 9.37 points (10-point scale). The first version of Russian synthesis rated averagely 5.58 points, while the improved synthesis – only 4.96 points. Though the difference in evaluation of comprehensibility of both syntheses is not statistically significant, the lower evaluation rate received by the improved version of synthesis is a bit unexpected. A more detailed analysis shows that an exceptionally low evaluation of the improved synthesis is given by the subjects with a long duration of work with the first version of synthesis.

6. For the Lithuanian subjects highest rating was given to human voice – 8.96 points. The first synthesis rates 5.27 points, and the improved one – 5.71 points. Thus, the improved synthesis is more acceptable to subjects. The old synthesis is viewed by subjects as a rather decent instance of a robot's speech, while the new synthesis – as a poor variant of human speech.

7. Some 42 % of subjects who often use synthesis considered themselves, subjectively, to be synthetic speech users. Some 25 % of subjects tried synthesis just to satisfy their curiosity. 33 % of them did not come in touch with it. Thus we can see that more than a half of investigated LOGOS employees have something to do with synthesis. This substantiates the subjects' belief

that in future the frequcncy and intensity of the usc of synthesis will grow. This fact indicates that speech synthesis is regarded as a progressive and innovative thing.

8. Only 41.7 % of respondents had formed a true notion of the concept of "structured digitized talking book". The rest 58.3 % had a vague idea about it. The same proportion emerged from the answers to the questions about the main parts of digitised book, i. e. its chapters and paragraphs. Footnotes and picture descriptions (25 % positive answers) and hyperlinks (16.7 %) were even less clear to the respondents.

9. Almost a half of subjects (41.7 %) view their chances of mastering the reading program optimistically; they think they can perform any service after several hours of training. Only 16.6 % of subjects think that more time and more efforts are needed to master the reading program. Almost a half of subjects (41.7 %) say that they take delight in using the book. Only 8.3 % disliked the book, and 50 % were undecided.

REFERENCES

Blenkhorn P. Producing a text-to-speech synthesizer for use by blind people // Extra-ordinary human-computer interaction. Cambridge: Cambridge University Press, 1995. P. 307–314.

Duffy S. A., Pisoni D. B. Comprehension of synthetic speech produced by rule: A review and theoretical interpretation // Language and Speech. 1992, vol. 35, No 4, p. 351–389.

Gorenflo C. W., Gorenflo D. W., Santer S. A. Effects of synthetic voice output on attitudes toward the augmented communicator // J. of Speech and Hear. Res. 1994, vol. 37, No 1, p. 64–68.

Grumadienė L., Žilinskienė V. Dažninis dabartinės rašomosios lietuvių kalbos žodynas. Vilnius. 1997. P. 398.

Higginbotham D. J., Drazek A. L., Kowarsky K., Scally C. A. Discourse comprehension of synthetic speech delivered at normal and slow presentation rates // AAC: Augmentative and Alternative Communication. 1994, vol. 10, No 3, p. 191–202.

Higginbotham D. J., Scally C. A., Lundy D. C., Kowarsky K. Discourse comprehension of synthetic speech across 3 augmentative and alternative communication (AAC) output methods // J. of Speech and Hear. Res. 1995, vol. 38, No 4, p. 889–901.

Hou Z. Z., Pavlovic C. V. Effects of temporal smearing on temporal resolution, frequency-selectivity and speech-intelligibility // J. Acoust. Soc. Am. 1994, vol. 96, No 3, p. 1325–1340.

Humes L. E., Nelson K. J., Pisoni D. B., Scott E. Effects of age on serial recall of natural and synthetic speech // J. of Speech and Hear. Res. 1993, vol. 36, No 3, p. 634–639.

Kasuya H. Significance of suitability assessment in speech synthesis applications // IEICE transactions on fundamentals of electronics communications and computer sciences. 1993, vol. 76, No 11, p. 1893–1897.

Keller E. Fundamentals of phonetic sequence // Fundamentals of speech synthesis and speech recognition. Basic concepts, state of art and future challenges. Christer et al: John Wiley and Sons, 1995. P. 5–22.

Koul R. K., Allen G. D. Segmental intelligibility and speech interference thresholds of high-quality synthetic speech in presence of noise // J. of Speech and Hear. Res. 1993, vol. 36, No 4, p. 790–798.

Marcus ST. M., Syrdal A. C. Speech: articulatory, linguistic, acoustic and perceptual descriptions // Applied speech technology. Boca et al.: CRS Press, 1995. P. 1–41.

McNaughton D., Fallon K., Tod J., Weiher F. Effect of repeated listening experiences on the intelligibility of synthesised speech // AAC: Augmentative and Alternative Communication. 1994, vol. 10, No 3, p. 161–168.

Murray I. R., Arnott J. L. Toward the simulation of emotion in synthetic speech – a review of the literature on human vocal emotion // J. Acoust. Soc. Am. 1993, vol. 93, No 2, p. 1097–1108.

Murray I. R., Arnott J. L. Implementation and testing of a system for producing emotion-by-rule in synthetic speech // Speech Communication. 1995, vol. 16, No 4, p. 369–390.

Osberger M. J., Maso M., Sam L. K. Speech-intelligibility of children with cochlear implants, tactile aids, or hearing-aids // J. Speech and Hear. Res. 1993, vol. 36, No 1, p. 186–203.

Pavlovic C. H., Rossi M., Eepesser R. Use of the magnitude estimation tcchnique for assessing thc pcrformance of text-to-speech synthesis systems // J. Acoust. Soc. Am. 1990, vol. 87, No 1, p. 373–382.

Paris C. R., Gilson R. D., Thomas M. H., Silver N. C. Effect of synthetic voice intelligibility on speech comprehension // Human Factors. 1995, vol. 37, No 2, p. 335–340.

Payton K. L., Uchanski R. M., Braida L. D. Intelligibility of conversational and clear speech in noise and reverberation for listeners with normal and impaired hearing // J. Acoust. Soc. Am. 1994, vol. 95, No 3, p. 1581–1592.

Preminger J. E., Vantasell D. J. Quantifying the relation between speech quality and speech intelligibility // J. Speech and Hear. Res. 1995, vol. 38, No 3, p. 714–725.

Preminger J. E., Vantasell D. J. Measurement of speech quality as a tool to optimise the fitting of a hearing-aid // J. of Speech and Hear. Res. 1995, vol. 38, No 3, p. 726–736.

Rahim M. G. Artificial neural networks for speech analysis / synthesis. London et al.: Chapman and Hall, 1994. P. 199.

Ralston J. V., Pisoni D. B., Mullennix J. W. Perception and comprehension of speech // Applied speech technology. Boca et al.: CRS Press, 1995. P. 235–281.

Rounsefell S., Zucker S. H., Roberts T. G. Effects of listener training on intelligibility of augmentative and alternative speech in the secondary classroom // Education and Training in Mental Ret. 1993, vol. 10, No 4, p. 296–308.

Santon F., Marchioni A., Susini P. Speech-intelligibility in the presence of an echo and noise // J. de physique. 1994, vol. 4, No 5, p. 537–540.

Schmidt-Nielsen A. Intelligibility and acceptability testing for speech technology // Applied speech technology. Boca et al: CRS Press, 1995. P. 195–231.

Smither J. A. Short-term-memory demands in processing synthetic speech by old and young-adults // Behaviour and Information Technology. 1993, vol. 12, No 6, p. 330–335.

Tucker P., Jones D. Voice as interface: An overview // Int. J. of Human-Computer Interact. 1991, vol. 3, No 2, p. 145–170.

Venkatagiri H. S. Effect of sentence length and exposure on the intelligibility of synthesised speech // AAC: Augmentative and Alternative Communication. 1994, vol. 10, No. 2, p. 96–104.

Werner S., Keller E. Prosodic aspects of speech synthesis and speech recognition // Basic concepts, state of art and future challenges. Chichester et al: John Wiley and Sons, 1994. P. 23–40.

SINTETINĖS ŠNEKOS KOKYBĖS VERTINIMAS: KELIŲ KOMPIUTERINIŲ SINTEZATORIŲ LYGINAMASIS TYRIMAS

Albinas Bagdonas, Feliksas Laugalys

Santrauka

Straipsnyje pateikiami kelių versijų lietuviškos ir rusiškos sintetinės šnekos suprantamumo ir lietuviškos, rusiškos, vengriškos bei itališkos sintetinių šnekų patrauklumo duomenys. Lietuvių ir rusų diktorių kalba yra suprantamesnė nei atitinkama sintetinė. Ankstesnė rusiškos šnekos sintezė blogesnė nei lietuviška ar patobulinta rusiška sintezė (PRS). Pagal sintetinamų garsų charakteristikas aiškėja dvi priešingos PRS tendencijos – pagal bendrą atpažinimo klaidų mažėjimą ji artėja prie natūralios šnekos, tačiau pagal klaidų homogeniškumą nuo pastarosios tolsta. Kadangi pirmoji tendencija vyrauja, bendra atstojamoji rodo PRS gerėjimą.

PRS suprantamumo ir patrauklumo koreliacija taip pat rodo jos didesnį artumą natūraliai šnekai. Tiriamiesiems PRS yra patrauklesnė nei ankstesnė rusiškos sintezės versija. Pastaroji, tiriamųjų nuomone, panašesnė į roboto šneką, o PRS – į blogą, tačiau jau žmogaus šnekos versiją.

Pagal patrauklumo duomenis natūralią šneką labiausiai vertina vengrų klausytojai, o kritiškiausi jos atžvilgiu yra italai. Visos tirtos sintetinių šnekų versijos vertinamos kaip mažiau patrauklios nei natūrali šneka, tačiau jas patobulinus šis vertinimas švelnėja.

Įteikta 2002-03-14