DOCUMENT RESUME

ED 041 260	48	AL 002 461

AUTHOR	Scholes, Robert J.
TITLE	Perception of Synthetically Created Consonants by Speakers of Various Languages. Final Report.
INSTITUTION	Florida Univ., Gainesville.
SPONS AGENCY	Institute of International Studies (DHEW/OE), Washington, D.C.
BUREAU NO	BR-8-D-041
PUB DATE	15 Dec 69
GRANT	OEG-4-8-080041-0051-014
NOTE	13p.
EDRS PRICE	EDRS Price MF-$0.25 HC-$0.75
DESCRIPTORS	*Artificial Speech, *Consonants, Language Research, *Languages, *Phonetics, Phonology, *Recognition, Syllables

ABSTRACT

In this research, synthesized speech sounds were presented to native speakers of various languages. The sounds were intended to systematically explore the front-to-back place of articulation dimension while holding voice and manner constant. Five types of initial consonants were investigated in these studies: voiced stops, voiceless stops, voiceless (weak) fricatives, nasals, and liquid/semi-vowels. They were subjected to evaluation for "naturalness" by trained phoneticians and were presented to native speakers in an example word test. The general intelligibility and naturalness of the synthesized materials proved to be inadequate for the tasks in mind. As a consequence, no detailed results are given, and a re-emphasis on the research utility of speech synthesis systems is called for. (Author/JD)

FINAL REPORT

Project No. 8-D-041

Grant No. OEG-4-8-080041-0051-014

PERCEPTION OF SYNTHETICALLY CREATED CONSONANTS BY SPEAKERS OF VARIOUS LANGUAGES

December 15, 1969

U.S. DEPARTMENT OF
HEALTH, EDUCATION, AND WELFARE

Office of Education
Bureau of Research

FINAL REPORT

Project No. 8-D-041

Grant No. OEG-4-8-080041-0051-014

Perception of Synthetically Created Consonants by Speakers of Various Languages

Robert J. Scholes

University of Florida

Gainesville, Florida 32601

December 15, 1969

The research reported herein was performed pursuant to a grant with the Office of Education, U.S. Department of Health, Education, and Welfare. Contractors undertaking such projects under Government sponsorship are encouraged to express freely their professional judgment in the conduct of the project. Points of view or opinions stated do not, therefore, necessarily represent Office of Education position or policy.

U.S. DEPARTMENT OF
HEALTH, EDUCATION, AND WELFARE

Office of Education
Bureau of Research

Acknowledgements

A number of people aided in the research described in this report. Miss Anne Morse ran subjects and helped prepare tapes for subjects. Cecil Coker and Peter B. Denes of the Bell Telephone Research Laboratories allowed me to use their synthesis set-ups and taught me how to use them.

Summary

In this research, synthesized speech sounds are presented to native speakers of various languages. The sounds are intended to systematically explore the front-to-back place of articulation dimension while holding voice and manner constant. Each subject will, due to his specific native language, find only some of the potential places of articulation relevant to his own language. He is asked to label those which he does find to be relevant by providing a word which illustrates the sound. In addition, certain sounds will be assigned to one category by a speaker of one language, but to another category by a speaker of a different language. While the number of different labels provided for a given set of consonantal stimuli provide an objective means of determining the phonemic distinctions for those stimuli, the variations across languages in the categories to which specific stimuli are assigned provide an objective means of typologizing and comparing phonemic systems.

Five types of initial consonants were investigated in these studies: voiced stops, voiceless stops, voiceless (weak) fricatives, nasals, and liquids/semi-vowels. They were subjected to evaluation for 'naturalness' by trained phoneticians and were presented to native speakers of 20 different languages in the example word test.

The general intelligibility and naturalness of the synthesized materials proved to be inadequate for the tasks in mind. As a consequence, no detailed results are given, and a re-emphasis on the research utility of speech synthesis systems is called for.

Introduction

The purpose of this work is to develop a procedure which will:

1) make the process of phonemicization an objective task.
2) make the process of phonemicization a short task.
3) equate the phonemes of a language with the perception of the users of that language.
4) make the task of typologizing languages on the basis of phoneme patterns objective.

Within the framework of our research these goals are not separable --although they might well be in some other approach. Due to their close linkage, we will focus attention on the process of phonemicization itself.

At present, phonemicizing a language is neither short nor objective. The standard procedure may be outlined as follows:

a) the analyst gathers a large corpus of utterances of the object language
b) the corpus is transcribed--i.e., written in some form of a phonetic notation
c) the transcribed corpus is scanned for cases of phonetically similar sounds in complementary distribution.

Although much work follows Step c), this will be sufficient to demonstrate our point.

The gathering and transcribing of the sufficiently large corpus typically takes several months--the analysis may occupy several more months. The procedure we will propose takes at most a few hours.

The biases of the analyst come in both in the transcribing of the corpus and in its analysis. The phenomenon of hearing a foreign language as merely a distorted form of one's own is too well known to bear documentation here, see Scholes (1967b). Not only is the analyst's native language a cause of "interference" in his hearing of a different language, the same linguistic background will dictate in large part the set of pseudo-phonetic symbols with which he will transcribe it (e.g., note the Indo-European biases in the International Phonetic Alphabet).

The major deterents to objectivity and explicitness are, however, in Step c) and involve the notions "phonetically similar" and "complementary distribution." Not only has the obvious question as to just how similar two sounds have to be in order to be considered with respect to distribution never been answered, the more basic question of what is meant by the term at all is largely unexplored. As for complementary distribution, any two sounds can be shown to be complementary distribution if the context is extended far enough. For example, given a sound x and a sound y which are "phonetically similar" where x always occurs in the context A_B and y never occurs in the context A_B, we then say that x and y are allophones (variants) of a single phoneme; if, however, both x and y occur in the context X_B (where X may be anything, including A) we say that they are different phonemes since they contrast in the environment X_B. So long as the notion of context with respect to distribution is not defined, there is no way of determining which statement regarding x and y is true and which false.

Such problems do not arise at all in the procedure proposed herein.

The procedure we have in mind is as follows:

1. Tape recordings of synthetic speech sounds are made. These tapes contain a sufficient number and range of stimuli to exhaust the possible phonetic bases for the phonemic systems of the languages of the world.
2. These tapes are presented to native speakers of languages under investigation. The informant either responds to each stimulus by saying whether or not it sounds like one of the speech sounds of his language and, if it does, which sound or by saying whether a pair of sounds are the same or different.
3. His responses are plotted against acoustic mappings of the stimuli to determine the number and types of phonemes in the informant's language.

The first attempts to develop an automatic phonemicization procedure were directed toward simple vowel sounds. A set of vowels which covered the general range of vowel qualities found in speech were synthesized on the IBM (San Jose, California) speech synthesizer. These synthetic vowels were categorized by over 200 speakers of over 30 languages. Their categorizations yielded phonemic vowel inventories which were: 1) compatible with accepted phonemicizations, 2) statable in terms of category (phoneme) and range (phonetic variants) analysis, 3) comparable through the physical specifications of the stimuli, 4) derived in a few minutes for each speaker. These experiments and results have been reported in a number of publications: Scholes (1967a, 1967b, 1968a, 1968b, 1968c). A paper on this work was presented at the Linguistic Society of America meetings in December, 1965.

During the period February 1, 1967 to January 31, 1968, these techniques and goals had been applied to selected consonantal stimuli under the sponsorship of the office of Education (OEC 2-7-068486-2677). In this research, synthesized voiceless stops and voiceless weak and strong fricatives were generated and tested in open-ended categorization tasks. The synthesis was done at Haskins Laboratories in New York City and the testing at the Communication Sciences Laboratory at Florida. Consonantal stimuli of a single class (e.g., voiceless initial stops) were generated in a /#______a/ frame and presented in pairs for same:different judgments by speakers of various languages. It has been shown (by Liberman and others of the Haskins staff) that a speaker's ability to discriminate between speech-like stimuli is very sharp when the stimuli are members of different phonemes and very poor when the stimuli are members of the same phoneme; Liberman, et al,(1957). Consequently, when a set of voiceless stops which covers the entire front to back range of articulation is presented to a speaker of some language in which there are n voiceless stop phonemes, his ability to detect differences should increase sharply as the phonetic borders between the various phonemes are encountered. There should, then, be n peaks in his judgments of differences. Preliminary work on these consonants indicates that this is true. Abramson and Lisker's (1964) work on the voice:voiceless distinction in stops also substantiates the hypothesis.

The work done during the period of this grant (7/1/68 through 10/31/69) involved synthesis and testing of sets of voiced stops, voiceless stops, voiceless fricatives, nasals, and semi-vowels.

Using the speech synthesis system of Peter Denes of the Bell Telephone Laboratories, several hundred stimuli were generated. These are all of the form CV where V is a constant /a/ - type vowel, and C is a voiced stop, voiceless stop, voiceless fricative, nasal, or semi-vowel. For each such class of consonantal onset, formant transition configurations are manipulated in an attempt to cover the articulatory front-to-back range. The full set of stimuli have been subjected to judgment by trained phoneticians and those sounds which were judged to be un-speechlike have been eliminated. Randomizations were then prepared of the speechlike stimuli, and they are being presented to speakers of various languages in an attempt to determine the acceptability of these sounds for the perceptual categorization task.

Finding native speakers of other languages proved to be a difficult task on the University of Florida campus. One part-time research assistant spent many fruitless hours waiting for appointments to show up, talking to various foreign student associations, etc. For these reasons, copies of the tapes and sets of instructions and answer sheets were sent to several friends who had indicated a willingness to run the tests on their foreign student populations. However, this approach also proved to be unusable since results were not sent to me and tests were, apparently, not run. Finally, a fair number of subjects were run during the summer of 1969 by Miss Anne Morse.

Methods

During the period December 2nd through December 6th, 1968, the author used the speech synthesis system of Dr. Peter Denes of the Bell Telephone Research Laboratories. This system involves a terminal analogue synthesizer controlled via a DDP-224 computer. The programming required of the user is quite simple and easy to learn. One specifies, for each of 12 parameters, some beginning value, an end value, and a time domain. For example, for the first formant transition of an initial stop, the specifications might be: initial value; 250 cps, end value; 500 cps, time domain, 50 msecs.

This system was used to generate sets of CV sequences where the vowel is held constant (an a-type) and the initial consonant is a voiced or voiceless stop, a voiceless fricative, a nasal, or a semi-vowel. For each such consonant type, a basic pattern giving the correct voicing and manner perception is constructed. Within each such basic pattern, 2nd and 3rd formant initial values are varied systematically so as to explore the full range of front to back places of articulation for the given class.

For the voiceless stops, all parameters are held constant except F2 and F3 (second and third formants). F3 is given one of three values: 1536, 2500, or 3584 cps at onset; F2 has starting values of: 512, 717, 922, 1127, 1332, 1537, 1742, 1947, 2152, 2357, and 2560 cps. The possible combinations of these initial F2 and F3 values result in 26 sounds.

These same F2 and F3 onset values were used for the other classes of consonants; other parameters being modified to produce the correct voice and manner specifications.

In the first testing all of the synthesized CVs were randomized in a single list and presented to trained listeners in the Communication Sciences Laboratory. These subjects were asked simply to transcribe the initial consonants, where appropriate, and to indicate which sounds seemed inappropriate for transcription - that is, which sounds did not resemble speech sounds of any type. On the basis of these judgments, the original list of 156 stimuli (there were two sets of 26 voiceless stops, 26 voiced stops, 26 voiceless fricatives, 26 nasals, and 26 semi-vowels) was cut to 103 speechlike sounds.

The 103 CVs judged to be speech-like were then rerandomized onto a single tape for further testing. In these tests, the stimuli were presented to speakers of various languages who were asked to write down, for each CV which seemed to them to be a CV occurring in their language, some example word. For example, a speaker of Russian, hearing a stimulus such as /da/ might write down дать 'give'. The languages investigated in this manner were: Japanese, Hindi, Russian, Yugoslavian, Tagalog, German, Polish, Hungarian, Thai, Iranian, Telugu, Czech, Indonesian, Chinese, Turkish, Hebrew, French, Sinhalese, Arabic, and Danish.

Results and Findings

Analysis of these materials has proceeded just far enough to indicate that further investigation is unquestionably a waste of time. By this, I do not mean to imply that the research goals and techniques are not tenable, but only that the stimuli which were used in my tests are not good enough. Since what we wish to look at is how the place of articulation varies across languages for a single set of consonantal stimuli, we require that the voice and manner dimensions be held constant within each such set. For our stimuli, this does not happen. For example, a stimulus CV where the consonant is intended to be a voiced alveolar stop is variously heard as: /d/, /t/, /y/, and /l/. Although this example may suggest that all subjects are hearing the stimulus as being of the same place and that their native language backgrounds account for the difference in voice and manner perception, other examples can be found to thwart this hope; say, this set of categorizations from a stimulus of the same (intended) set as the one above - /b/, /l/, /r/, /y/.

As a consequence of the fact that unpatterned responses such as those shown in the examples are the general rule for all of the synthesized materials, it appears fruitless to pursue further analysis of these materials.

Lest the reader of this report regard the above as an admission of personal incompetence, I should like to claim that the problem in this research is not to be found in the techniques, goals, or execution, but rather in the synthesis itself. I would claim two deficiencies of synthesizers which must be rectified before research such as that pursued here can be confidently carried out. One deficiency is the synthesizer's overall intelligibility. Although whole sentences are quite easy to understand when generated by rule (by programming only) (as demonstrated by any number of research groups), gaining intelligibility for nonsense syllables is a much thornier problem. As we know, intelligibility for humanly produced CVs is not nearly 100% (in general, intelligibility decreases with decreasing context). Thus, if we take away some of the acoustic cues which contribute to intelligibility, the scores will drop even more, and we may be asking, in our study, for subjects to label stimuli which they simply cannot identify.

The second deficiency of synthesis systems at present is their language-specific range. Most synthesis systems are designed to be able to produce just those speech sounds found in English (although some include Swedish, French, or Japanese in their repetoires). No synthesizer which I have heard of has even the capability to eventually be programmed to produced such sounds as ingressive clicks or pharyngeal spirants. Consequently, attempted to use a device built to speak English to generate, say, Arabic, is at best a trial and error procedure.

In summary, then, it must be claimed that the research interests of linguists and phoneticians are not yet handleable by synthesis.

Although this research must in all honesty be said to be a loss, it is hoped that my failure will prompt others to think of synthesis less as a technique for generating acceptable instances of English utterances and more as a technique for doing controlled linguistic research.

Table 1 is given below as an illustration of the kind of responses given for a particular set of stimuli.

	Tagalog	German	Polish	Hungarian	Czech	Indonesian	Turkish	Hebrew	French
1	r	r	r	r	r	d	r	r	r
2	d	-	dz	d	l	b	d	-	d
3	l	-	y	y	y	t	d	y	d
4	-	-	t	-	d	-	u	-	d
5	y	y	l	y	y	y	y	y	y
6	w	v	-	v	v	w	l	-	w
7	w	-	-	v	v	w	y	-	w
8	-	r	-	r	l	-	u	-	l
9	y	l	-	y	-	y	l	-	-
10	r	y	-	-	b	l	r	-	-
11	y	v	-	-	y	y	r	y	y
12	-	-	-	r		w	-	r	-
13	l	l	-	v	r	l	-	-	l
14	y	w	-	y	v	d	l	-	-
15	l	v	l	r	d	l	y	-	l
16	y	-	r	y	d	w	-	y	-
17	w	-	l	w	-	w	-	r	l
18	w	-	l	w	r	-	-	-	l
19	-	-	l	y	l	l	-	l	-
20	w	l	y	r	d	y	-	l	l
21	l	-	l	y	-	-	v	r	-
22	y	y	l	y	-	w	v	-	-
23	-	-	l	r	-	w	d	l	l
24	w	v	l	w	r	l	v	-	-
25	l	l	l	y	d	w	l	y	-
26	-	-	y	w	r	y	y	r	d

Table I. Some cross-language responses to one set of 26 stimuli.

Conclusions and Recommendations

The only feasible conclusion to the efforts to use synthetic CVs to investigate the native-language determined categorizations of acoustic stimuli is that this is not a workable project at the present state of speech synthesis.

Synthesizers must be improved with respect to naturalness and intelligibility and must be constructed with the whole range of articulatory possibilities in mind before such work can be carried out.

One line of research currently being pursued which at least provides a hope for a general purpose research synthesizer is the work on articulatory analogue synthesis. Cecil Coker of the Bell Telephone

Laboratories, for example, is working on such a model; where the input specifications are essentially articulatory 'target' positions in sequences. The transition from one such target to the next is done via internal computations which are analogues of vocal tract masses and speeds. Although this research, in terms of the articulatory-analogue programming is entirely consistent with the requirements for truly research oriented synthesis; Coker's goals are still limited by the nature of the synthesizer which produces the sounds themselves. Even were he able to completely program the articulation of the vocal tract, the synthesizer could not manifest many of the possible (and common) configurations of that tract.

References

Scholes, Robert J. (1967a) Phoneme categorization of synthetic vocalic stimuli by speakers of Japanese, Spanish, Persian, and American English, Language and Speech, 10.1,

(1967b) The categorization of synthetic speech sounds as a predictive device in language teaching, Journal of English as a Second Language, 11.2,

(1968a) Categorial responses to synthetic vocalic stimuli, Language and Speech, 11.2,

(1968b) Perceptual categorization of synthetic vocalic stimuli as a tool in dialectology and typology, Language and Speech, 11.4,

(1968c) Derivation of phoneme inventories by native speaker responses to synthetic stimuli, Final Report, Contract No. OEC2-7-068486-2677, Office of Education, Washington, D.C.

Abramson, Arthur S., and Leigh Lisker, A cross-language study of voicing in initial stops, Word, 20.3 (1964)

Liberman, Alvin, et al., The discrimination of speech sounds within and across phoneme boundaries, Journal of Experimental Psychology (1957)

APPENDIX D.--ERIC REPORT RESUME

Department of Health Education and Welfare
Office of Education

Eric Accession No.   ERIC REPORT RESUME

Clearinghouse Accession No.:	Resume Date: 12/17/69	IS DOCUMENT COPYRIGHTED? Yes [✓ No] ERIC REPRODUCTION RELEASE [✓ Yes] No

Title:
Final Report: Perception of Synthetically Created Consonants by Speakers of Various Languages

Personal Author(s):
Scholes, Robert J.

Institution Source:
University of Florida, Gainesville, Fla., Department of Speech

Report/Series No.

Other Source: | Source Copy

Other Report No.

Other Source: | Source Copy

Other Report No.

Pub'l Date: 17 Dec. 69 | Contract Grant No.: OEG-4-8-080041-0051-014

Pagination, etc.:

Retrieval Terms:
Speech Synthesis
Phonemicization

Identifiers:

Abstract:

Consonant + vowel syllables are generated by speech synthesis such that the vowel is a constant a-type and the consonants vary systematically in initial second and third formant values within consonant types. Voice and manner types of consonants were stops (voiced and voiceless), voiceless fricatives, nasals, and liquids and semi-vowels. These CVs are then presented to native speakers of various languages who are asked to provide example words for the stimuli which are found to be relevant to their language. No specific results are given due to the fact that the synthesized materials were not adequately intelligible.