Heuristic models for automatic recognition of spoken words

Derkach, M.

journal: STL-QPSR
volume: 11
number: 1
year: 1970
pages: 039-049

http://www.speech.kth.se/qpsr
C. HEURISTIC MODELS FOR AUTOMATIC RECOGNITION OF SPOKEN WORDS

M. Derkach

General

Investigations of speech perception and recognition were started in 1963 in the biophysics group of the Department of Physiology of Lvov University. The problem of natural human speech perception is studied with the view of searching for possibilities of developing an adaptive automatic system based on a computer which could recognize speech messages of a restricted number of words and simple phrases independently of the manner of the speaker's pronunciation.

Attempts of using heuristic procedures in recognizing algorithms have been made and various kinds of interaction between directly perceived information from the speech signal and a priori linguistic information are foreseen. This process is based on a continuous synthesis and test of possible words (lexical hypotheses) from a running phoneme description of the message first in a preliminary generalized form and then ascertaining some of them by searching for additional arguments and distinctive features of the phonemic and phonetic structure of the message. Different kinds of feed-back loops providing interaction between various levels of the whole hierarchical system are incorporated.

An active control of the whole procedure of extracting useful information is provided by the highest level, namely by the lexical one and is based on the analysis of the spoken message according to the linguistic properties and rules of the given language.

Our general approach to the problem is based on the philosophy that the natural processes of speech perception as well as automatic speech recognition models should include some general rules of mental activity of the receiver ensuring a meaningful contact between the speaker and the listener in spite of frequent omissions of distinctive features and phonemes in the message.

Speech perception incorporates a complex strategy of extracting necessary features from the message in due shape, time, place, and volume for maintaining the contact between the participants. An adequate model of the natural processes should therefore have enough

* Guest researcher from Lvov University, USSR, from February 17, 1970.
degrees of freedom to be independent of the shortcomings of the speech wave produced by an individual speaker as well as of the variability of speech and hearing and phonetic training of the listener. This approach is close to the general philosophy of the problem of speech perception and recognition outlined by Fant (8, 9, 10).

Till 1970 two categories of manner of speech production have been taken into consideration - vowels and voiceless fricatives which alternate forming words like .... VCVC .... The recognizing material was restricted to words pronounced in isolation which did not form sentences or phrases. They were represented by real Russian words, infinitives of verbs and nominative singular of nouns, selected from the Russian vocabulary. Some of the words terminated on phonemes not belonging to the inventory. In this case the end of the word was cut off and was not taken into further consideration.

The list of words prepared for recognition procedures is shown in Table III-C-I. It should be noticed that some of the words have very small differences in their phonemic structure. They are marked in Table III-C-I. Stressed vowels are underlined. Eight standards of vowels and the same number of fricatives were used, namely

[u] [o] [a] [a'] [e] [e'] [y] [i];
[s] [s'] [f] [f'] [ch] [ch'] [sh] [sh']

(the mark [''] means palatalization of the phoneme).

The main task of the investigations consisted in developing a system for recognition of the list of the above-mentioned 73 words. The system was not adopted to these words or to the speakers which pronounced them. According to the formulated task the main direction of the study was concentrated on the problems of segmentation of the spoken message, phoneme identification of its elements, and the overall organization of the recognition system.

Principles of segmentation of spoken words

Two main approaches to the segmentation were studied. The first one consists in detecting boundaries on the time axis of the speech wave which would produce a number of segments greater than the number of phonemes in the message (7). The segmentation precedes the phoneme identification of the utterance in this case. The type of
<table>
<thead>
<tr>
<th>USY</th>
<th>USA Sh'</th>
<th>USY Ch A</th>
</tr>
</thead>
<tbody>
<tr>
<td>U sh I</td>
<td>US' E Sh'</td>
<td>USY Ch A</td>
</tr>
<tr>
<td>U Ch O</td>
<td>US' E Sh'</td>
<td>ISUSh A</td>
</tr>
<tr>
<td>U Ch A</td>
<td>USO Ch</td>
<td>OSU Sh A</td>
</tr>
<tr>
<td>USA</td>
<td>OS' ES'</td>
<td>OSU Sh A'</td>
</tr>
<tr>
<td>OSO</td>
<td>OF' IS</td>
<td>EShAFO</td>
</tr>
<tr>
<td>OS' O</td>
<td>EP' ES</td>
<td>AF' ISh A</td>
</tr>
<tr>
<td>OSA</td>
<td>ISUS</td>
<td>FAShYS</td>
</tr>
<tr>
<td>OSY</td>
<td>SUCH A</td>
<td>ShAShYS</td>
</tr>
<tr>
<td>OS' E</td>
<td>SUSH A</td>
<td>SOF' IS</td>
</tr>
<tr>
<td>OS' I</td>
<td>SACh A</td>
<td>SUCHOS</td>
</tr>
<tr>
<td>QCh A</td>
<td>SACh A</td>
<td></td>
</tr>
<tr>
<td>ASY</td>
<td>SOCh A</td>
<td></td>
</tr>
<tr>
<td>ACh A</td>
<td>SOFA</td>
<td></td>
</tr>
<tr>
<td>Ash U</td>
<td>SO' A</td>
<td></td>
</tr>
<tr>
<td>AS' A</td>
<td>SOS' E</td>
<td></td>
</tr>
<tr>
<td>ESE</td>
<td>SOSA</td>
<td></td>
</tr>
<tr>
<td>EF' I</td>
<td>SOSO</td>
<td></td>
</tr>
<tr>
<td>ECh O</td>
<td>SOSU</td>
<td></td>
</tr>
<tr>
<td>ISh A</td>
<td>SOF' I</td>
<td></td>
</tr>
<tr>
<td>SYS</td>
<td>SAFO</td>
<td></td>
</tr>
<tr>
<td>S' ES'</td>
<td>SAF' A</td>
<td></td>
</tr>
<tr>
<td>FAS</td>
<td>S' IFO</td>
<td></td>
</tr>
<tr>
<td>F'I Sh</td>
<td>FASO</td>
<td></td>
</tr>
<tr>
<td>Sh Y Sh</td>
<td>FAS' E</td>
<td></td>
</tr>
<tr>
<td>Sh Y F</td>
<td>FASA</td>
<td></td>
</tr>
<tr>
<td>Sh E F</td>
<td>SYCh A</td>
<td></td>
</tr>
<tr>
<td>Sh E S</td>
<td>ChOChO</td>
<td></td>
</tr>
<tr>
<td>Sh A Ch</td>
<td>ShOSE</td>
<td></td>
</tr>
<tr>
<td>Ch US</td>
<td>ShAS' I</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ShOF' O</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ChASA</td>
<td></td>
</tr>
</tbody>
</table>

**TABLE III-C-I.**

Experimental vocabulary of recognizing words.
segmentation consists in making marks corresponding to significant reconstructions of the source (noise/voice) and the dynamics of articulation. The principal point here is to pick up not only sustainable sound segments but also to observe tendencies of changes of articulation in the spoken utterance.

The whole spectrum envelope was taken into consideration for this purpose. Spectral analysis of the speech material was made with a bandpass filter analyzer. A +6 dB/oct preemphasis was utilized. The audio range from 90 to 18000 Hz was divided into 23 one-third octave bands in this analyzer with 50 msec time-constant of smoothing of the time-envelope in each band. The output amplitudes of the filters were sampled every 32 msec.

The correlation function of neighbor spectral cuttings was taken as the base of segmentation procedure. Especially the extreme points of the rate of change of the correlation, $\Delta r$, between adjacent spectral cuttings in time proved to be quite successful for segmentation of utterances consisting of vowels and fricatives at least. The function of correlation speed, $\Delta r$, provides much better results in segmentation than the function of correlation between adjacent spectrum shapes $(r_n, n+1)$ itself, which can pick up only well expressed sustained parts of the sounds. Provided the frequency of spectrum sampling is kept constant it is counted in a simple way as

$$\Delta r_n = r_{n-1} - r_n, n+1.$$

An example of segmentation of the word "susha" (a dry land) is given in Fig. III-C-1. Extremal points of speed of the correlation $\Delta r$ are shown by points and boundaries of the segments are marked by arrows. The contour of the oscillogram is pictured at the bottom of the diagram.

It seems to be a special problem to solve how to find the boundaries of phonemes after obtaining the boundaries of the segments. Some linguistic rules have to be applied in this case which are incorporated in the structure of the language but not of the signal itself. Some restricted grammar rules have accordingly been adopted in our program of phoneme segmentation which does not allow a vowel to be followed by another vowel and a consonant by a consonant. The grammar prescribes sequences like .... CVCVC.... This means that, for instance,
Fig. III-C-1. Segmentation of the word "susha" by means of the correlation speed function.
Fig. III-C-2. Example of time normalization in the segmentation procedure.
when some diphthong happens to occur the program will join all vowel-like segments into the united one and then look for the representative description of it. This process is shown in Fig. III-C-2 where the word "afisha" (a playbill) is segmented. The segments of the same group (V or C) are joined together (line IV) and are then represented by the spectrum of the most sustainable part during this interval (line V). Lines I, II, and III are numbers of spectrum cuttings at distances 32 msec apart, the contour of the oscillogram and the curve of correlation speed, respectively. In other words, the time normalization of the number of segments, e.g. decreasing their numbers to the number of phonemes in the message demands the administering of phonological rules of grammar which were a priori used when the list of words for recognition was selected.

The second principle of segmentation is based on the close interaction between the segmentation and phoneme identification procedures in the time domain. The segmentation does not follow here as a result of an independent procedure which cuts the signal into pieces because of its physical inhomogenity. The segmentation of the message comes here as a result of evaluation of the possibility to find in the pronounced message the sequences of segments which might correspond to the possible words tested one by one. This approach which only recently started in the laboratory seems to be more convenient for further linguistic analysis of the message due to its close interaction with the phoneme interpretation of the message in time.

**Principles of phoneme identification of sounds and segments**

The main part of the experimental material is represented by vowels. Different functions of spectral descriptions and different kinds of evaluation functions for making phoneme decisions were tested. Most of them incorporated a statistical approach to the problem of phoneme decoding of the spectra as well as to the establishment of the most informative spectral areas for this aim. Different kinds of description of phoneme standards were tested. Some of them were based on the description of the whole shape of spectral envelope which generalizes the teaching material. In the other group of experiments an attempt was made to obtain a restricted number of more informative spectral regions and to use them in the recognition procedures.
One group of useful parameters reflected the spectral areas with minimal statistical dispersion within the teaching material. The energy of the output of one-third octave filters were normalized with respect to the overall energy of the sample

\[ X = (x_i), \quad x_i = \frac{E_i}{E_o}, \quad i = 1-23 \]

All possible bandwidths with all possible mean frequencies were looked through in the computer experiment and three non-overlapping spectral areas with minimal dispersion of intensity in the teaching material were chosen as the characteristic features of the phoneme (1). The 23-dimensional space of description of the phoneme was reduced to the three-dimensional one in this case.

About 90% of correct responses were obtained in an experiment with the automatic recognition of isolated vowels pronounced by the control speakers (2). 90% of correct perception was reached in the group experiment when isolated vowels were synthesized from a neutral vowel by means of amplifying the characteristic spectral areas and switching off all the others (3) (according to the specifications for each vowel).

Characteristic phoneme spectral areas were used in the experiments on automatic classification of vowels cut off from the spoken words. No trustworthy difference between using the whole spectral envelope and 3-5 characteristic areas was obtained. The effectiveness of automatic recognition of vowels taken from the teaching material was usually about 90% and that for a new, control material - about 70%. The general conclusion has been made on the basis of experimental studies that the effectiveness of recognition of the control material does not change considerably when either the phoneme standard description or the evaluation function are simplified in wide amounts (4). It allows the use of simple procedures for phoneme decoding, e.g. selecting any individual spectra of any speaker as the phoneme standard or to describe the statistical standards in a very approximate form as well as to administer simple evaluation functions for phoneme decisions making without any losses in the effectiveness of phoneme decoding the speech sounds.
The matrix-like presentation of the spectral envelope and phoneme standards was finally adopted \(^{5}\). It is shown in Fig. III-C-3. The matrix consists of 23 overlapping 2/3 octave bands in the range 90-18000 Hz with five 6 dB amplitude levels. The shape of the individual envelopes is marked in the matrix \([a_{j,k}]\) by units in its cells, as it is shown at the top of the figure. The phoneme standards are presented by the same matrices but each cell is given the numerical weight according to how frequently it was met in the teaching material (the matrix \([e_{j,k}]\) at the bottom of the figure. A three-level system of statistical evaluation is applied. The modal cells are given the highest mark 2, the most rarely occurring ones - the mark 0, and all others are marked by the indefinite mark 1. The simple sum of marks sampled by non-zero-marked cells of the recognizing sound on the phoneme standard matrices was found to be convenient as the evaluation function for phoneme decoding the sounds and representative segments in words.

The second basic conclusion in our study of phoneme decoding of spectra is concerned with the impossibility to make unanimous phoneme decisions for sounds and segments pronounced by new speakers and within new speech utterances. It follows from the statement that any teaching material of sounds generalized in the phoneme standard cannot be argumentally considered as a material which represents the whole general population of sounds of the given phoneme. Therefore, the phoneme decisions have to be made in the form of non-unanimous generalized phoneme codes, based on threshold principles. This approach to phoneme recognition which corresponds well with the general physiology of perception makes it possible to avoid mistakes and to continue the analysis of the speech message in the whole system using more appropriate features in feedback loops.

It is a frequent situation that only a few words may be constructed from the sequence of ambiguously decoded phoneme segments and the choice of one among them should be argumentally made. It has proved useful to change the space of spectral features in this case in order to concentrate on pairwise oppositions, i.e. distinctive differences. The frequency location of these so-called differential features were investigated by applying different statistical measures, especially \(\chi^2\).
Fig. III-C-3. Description of the spectrum envelope in the matrix form $[a_{j,k}]$ and that of the phoneme standard $[e_{j,k}]$. 
and t. Three non-overlapping parts of the spectrum with a maximum statistical distance were selected for each pair of phonemes which served as additional standards in the space of differential features. The choice of alternative feature sets in phoneme decoding is controlled by a higher level of the recognizing system. In specific, the differential features are administered after the lexical level has selected the competing words among the generalized phoneme sequences and the degree of ambiguity has been determined.

**Organization of the hierarchical model of speech perception in order to recognize isolatedly pronounced words**

The algorithm was proposed and the computer program was written for recognizing 73 Russian words consisting of alternating vowels and voiceless fricatives (8). The general scheme of the recognition model is shown in Fig. III-C-4. The analysis of the spoken words starts out from a parametric description converting the utterance into a dynamic spectrogram (0 level). The informative transformations come at the first level. They prepare the dynamic spectrogram for the next transformations, i.e. segmentation and phoneme decoding of segments using different kinds of features.

The phoneme descriptions of the segments are completed at the second level as a result of phoneme decoding. All the phoneme codes which lay above a certain threshold are taken into consideration.

Sequences of unanimous phoneme codes build word-like syllables at the third level of the system. Real words listed in the lexical vocabulary are separated from nonsense syllables and selected. Repeated cycles of analysis which correct and ascertain the phoneme structure of the utterance are controlled by the lexical level which is the highest one in the model.

More detailed structure of the recognizing algorithm is shown in Fig. III-C-5. Its main procedures consist of the following:

1. Spectral analysis of the spoken word \( f(t) \) and producing a sequence of spectral sections sampled every 32 msec (SA).
2. Tracing the dynamics of the overall intensity for analyzing the stress in the word (I).
3. Primary segmentation of the message by means of measuring the correlation between the neighboring spectral sections (PS). The extreme points of rate of spectral correlation are used as temporal boundaries of primary segments.
Fig. III-C-4. General scheme of the recognizing system.
Fig. III-C-5. Block scheme of the recognition algorithm.
Normalization of a number of primary segments in time and transformation of the message into a VCVC sequence using the rough distinctions in the spectral shape between vowels and fricatives (TN).

Selecting the representative spectral section within each phoneme of the sequence and administering the rules of their useful description (UD). Each phoneme appears then as a point in the multidimensional space of primary useful features X.

Phoneme decoding the representative segments (PH₁) using the threshold principles (η²) of phoneme decisions. The sequence of generalized phoneme codes (Rₚₜ) comes as a result.

Construction of all possible phoneme sequences (PhS) from a generalized phoneme code and selection of real words among them (VS) using the vocabulary (Voc).

When the single word is found in the vocabulary it is typed as a final result of recognition (W₀).

When a unanimous result is obtained (W₁) the program makes repeated analysis of the selected words using additional information about their phonological structure according to the possible phonemic deviations of pronunciation of the words stored in the vocabulary (AA). Some prosodical information (stress location) as well as some additional data about the intensity of the consonants are taken into consideration (AI) and all the evaluation marks are summarized (Σ).

The second loop of the feedback is foreseen when ambiguous decisions still remain (W₂). The phoneme contrasts in competing words are analyzed carefully under the vocabulary control (PhC). The space of description of contrasting segments is changed to the differential features description Y and the phoneme decision is made by means of the direct contrast of phonemes in pairs (PH₂). The process ends by typing out the recognized word.

The experimental examination of the program on the material of 73 words has shown 93% of perfect recognition of words pronounced by the male and female speakers which never had been used in the experiments. Only four mistakes were registered, namely,

- the word os'c was recognized as os'i
  - sofa as socha
  - shyf as shysh
  - isusha as isycha

Conclusion

We consider the algorithms and computer programs for the automatic speech recognition to be interesting not only for the practical aims of engineering but at the same time as fruitful methods of the

---

The threshold η is controlled by the results of stress analysis in the block I.
study of the natural process of speech perception itself. Different kinds of heuristic models based on the cybernetical principles of selforganization and selfcorrection are apparently perspective in this kind of studies.

Speech perception is an increasingly higher order organized cybernetic system which we can meet among the biological systems. The problem of organization of its aimful activity should be taken into consideration in the automatic speech recognition studies as well as the fact that the goals which the recognizing system has to attend are those of linguistic nature.

Fant (9) has made an analysis of the human behavior organization in the process of decoding the spectrograms of spoken messages. Many repeated evaluations and corrections of preliminary decisions are foreseen in this process and a powerful phonetical analysis is considered in this case. It seems to be useful to follow this strategy in automatic speech recognition computer programs.

Acknowledgments

I am indebted to Prof. G. Fant for fruitful discussions on contents and form of this paper. I am also grateful to Mrs. S. Felicetti for her expert editorial assistance.

References


refs. continue on next page


(10) Fant, G.: "Automatic Recognition and Speech Research", in this issue of STL-QPSR.