RECENT RESULTS IN SPEECH PROCESSING

By

G. Gordos

Institute of Communication Electronics, Technical University, Budapest

Received October 5, 1979

Presented by Prof. Dr. S. Csibi

1. Introduction

At the Institute of Communication Electronics, Faculty of Electrical Engineering, Technical University, Budapest, and its predecessor, the Department for Wire Telecommunications, investigations have been conducted on speech processing since the mid-sixties. A summary of the main directions of research is given here along with some results in the field of the diagnosis of twin-zygosity.

2. Short survey

Speech processing is the artificial realization of one or more functions of human speech, hearing and understanding, in short, the human speech-chain. Its main branches are:

— speech recognition
— speaker identification and
— speech production.

Speech recognition and speaker identification use methods related to those used in classification of infants' crying, which aims at determining the cause of the infants' cry (hunger, pain, etc.).

Some of the important applications of speech processing are:

— providing for a possibly two-way acoustic man–machine communication,
— typing after dictation,
— visualization of parts of speech for teaching deaf persons to speak,
— speech transmission with decreased channel requirements,
— ciphering,
— distinguishing between the mono- and binovularity of twins by means of acoustic tests for the determination of the hereditarity of certain illnesses (see under 4),
— diagnosis of deteriorations of speaking and hearing organs.

For better understanding the methods presented, it seems expedient to make a distinction between some principles of the interrelated fields of speech recognition and speaker identification.

The first essential result of speech recognition — the automatic detection of durably sounding vowels and certain voiced consonants — was based on the recognition of formant structures (see e.g. [3]). This result turned the attention to spectral investigations by means of linear filters. The essential element of the hearing organ model, proposed by HELMHOLTZ [1], anatomically proved and further developed by BÉKÉSI [2], is a set of loosely coupled resonators. This fact suggested for a time that by refining the filter system the most decisive phase of speech recognition could be accomplished. Though these investigations brought about some lasting results, the initial difficulties (uncertainty in the recognition of normally timed, instead of durably sounded vowels, the unmanageability of transitions modifying the meaning, dependence on the sex and on the person, etc.) have never been totally overcome. The principal limits of the method were formulated in the seventies (see e.g. [4]). At that time, the consequences of the Heisenberg's uncertainty relation, deduced for linear filters by DENNIS GÁBOR [5] and refined by HARKEVICH [6], and experimental data concerning the frequency-analyzing ability of the ear could be brought in relation [7]. The former ([5]) states that the frequency of a sinusoidal burst with a life-time of ΔT can only be estimated, by a system of linear filters, with an accuracy of Δf if $\Delta T \times \Delta f > C$, and the value of C is about 1 or 2, depending on the envelope of the sinusoidal burst. On the other hand; according to [7] the efficiency of the ear — in the frequency band of up to about 2 kHz — is described by the relationship $\Delta T \times \Delta f > 0.18$. Upon comparing this two values, one can conclude that, concerning the rate of spectral analysis, the ear cannot be modelled by a linear system of a filtering-type because it is by about one order of magnitude more efficient than the latter.

This is one of the causes that brought about a change in the strategy of speech recognition from the mid-'60-s. In addition to the computer simulation of the human hearing mechanism, speech recognition began to analyze sound waves in order to draw conclusions on the processes taking place in the speaking organs. It is obvious that the processes of speaking organs are determined by the message, and the aim of speech recognition is just to establish this message. To approach speech recognition from the side of speech formation promises advantages in a general sense too in that man's capability of producing acoustic phenomena is poorer than that of perceiving it. Thus, speech recognition can achieve equally valuable results by studying the mechanism of speech formation instead of the substantially more complex mechanism of speech perception.

A further important shift in the methods applied occurred due to a guess, becoming later a conviction (Klatt, Stevens, 1971), that even man is not capable of perfect recognition based on purely acoustic phenomena. That is why computerized speech recognition is operating on several levels with feed-back between them.

The subprocedures most often applied are:

- — procedure on acoustic level, acting by
  - — acoustic feature extraction
  - — classification and
  - — producing an acoustic element (phonem, apel, etc.),
- — procedure on linguistic level, which can be divided into
  - — word (lexical) level
  - — syntactic level and
  - — semantic level, and finally
- — procedure on pragmatic level, which deals with the speaking person, the significance of the communication, etc.

The procedures of higher levels control those of lower levels, choose from the possibilities offered by the latters, or instruct them to look for new possibilities.

3. A feature extraction system

In addition to a continuous survey of the state of the art in speech processing, the work carried on at this Institute embrace the acoustic level of speech recognition and the whole of speaker identification.

These investigations relied initially on several dedicated instruments and on computers like an Odra 1024 (up to 1974), later on a PDP 8 and at the end a PDP 11/40, but from 1978 the development of self-contained subsystems for microprocessor-aided feature extraction systems is also in progress.

As an illustration, the main characteristics of the feature extraction system based on the PDP 11/40 are enlisted. After sampling, with a variable frequency, and quantization, the sound wave is mapped into the background storage. By means of a shiftable window any of its segments can be visualized on a graphic or alphanumeric display for surveying. The window typically contains 512 samples. The samples neighbouring on the display can be generated by the "thinning" of the stored sample series, thus it is possible to obtain a bird's view image of the sound wave investigated (the time of a window varies between e.g. 40 ms and 1.3 s). To facilitate comparison of two parts of a wave, it is possible to display the two parts simultaneously. Two shiftable cursors provide for the selection of an arbitrary segment and for its definition for further processing. Using the two cursors time can also be measured (e.g. the duration of explosives).

The option of interpolation ensures the possibility for the transformation of the chosen segment to have sample number that fits further processing independently of the original number of samples in the segment. This permits e.g. the comparison of parts with identical phonetic values said with different speed (time fitting), the application of FFT, or the comparison of spectral functions to be discussed later (frequency fitting).

An essential possibility is that of modifying (erasing, replacing, amplitude modulating, etc.) the chosen segment. The primary aim of this possibility is the realization of indirect feature extraction. (The "feature" is that property of the sound wave which influences the recognition.) If, namely, a segment, which contains the feature, is erased or changed, the sound wave restored from the modified samples of the segment becomes unrecognisable. It is also used for working out hypotheses [8] on how much the reversion, in time, of the course of certain acoustic elements influences the recognition.

Spectral investigations are facilitated by fast Fourier transformation (FFT) and cepstrum calculation (in its every variation, such as cepstrum, clipstrum, middle-deprived clipstrum) performable on an arbitrary segment.

The transform, of a function $f(t)$,

$$c(q)=|F^{-1}\{\lg|F[g\{f(t)\}]|^{l}\}|^{l}$$

F denoting the Fourier transformation, in the case of $l=1$ is of "amplitude" character, in the case of $l=2$ is of a "power" character, and it is

a cepstrum if $g(x)=x$; a clipstrum if

$$g(x)=\begin{cases} +A, & \text{if } x\geqq 0 \\ -A, & \text{if } x<0 \end{cases};$$

and a middle-deprived clipstrum if

$$g(x)=\begin{cases} x-A, & \text{if } x\geqq A \\ 0, & \text{if } A>x\geqq -A \\ x+A, & \text{if } -A>x.) \end{cases}$$

A new feature is the dynamic cepstrum display, which successively depicts the cepstrums of segments, shifted by regular time intervals, of one or two time functions, thus displaying the changes of the cepstra like a motion picture.

Both FFT and cepstrum features are supplemented by routins for the interactive determination of the fundamental as well as formant frequencies.

Investigation of the segments in terms of auto- and cross-correlation is also possible.

The automatic determination of the average value and the standard deviation of the fundamental frequency (i.e. pitch) can also be performed. The procedure is based on the AMDF (Average Magnitude Difference Function), which is defined over a sample series $\{x(n)\}$, $n=0, 1, 2, \ldots N$, as follows:

$$y(k)=\frac{1}{L}\left[\sum_{n=l}^{l+L-1}|x(n)-x(n+k)|\right],$$

where k is the shift, $L<N$, $0\leqq l\leqq N-L+1$, and l is the starting point of the investigation. This function takes on zero, in theory, or a minimum, in praxis, at a shift K corresponding to the period, in time, if $\{x(n)\}$ is periodic.

In practice, the choise of Y contained in the relationships

$$k=K \text{ if } y(k)<Y$$

defining the minimum is very critical.

Our tests showed that satisfactory results could be obtained by choosing

$$Y_1=c\left\{\frac{1}{N+1}\sum_{n=0}^{N}\left\{x(n)-\frac{1}{N+1}\left[\sum_{i=0}^{N}x(i)\right]\right\}\right\},$$

$$c=0.6 \text{ to } 0.8$$

or

$$Y_2=\frac{1}{32}|\{x(n)\}|_{\max}$$

There are built-in provisions for keeping k within reasonable limits, and for automatically and equidistantly varying l. This ensures the recognition of all quasi-periodic parts of the sound wave and also the determination of its fundamental frequency.

Further features of the feature extraction and classification offered on one hand by the system based on the PDP 11/40 and on the other hand by the autonomous microprocessor subsystems already developed are performing the determination of the fundamental period, the recognition of explosives and the automatic measurement of their duration. These will be described in a later paper. (It is to be noted, that this subsystem proved to be useful also in a different research, namely in the implementation of an impulse-noise trap.)

4. Determination of twin-zygosity based on acoustic characteristics

The main aim of our work in the field of speech processing was the acoustic determination of twin-zygosity.

Establishing twin zygosity, i.e. distinguishing between monovular and binovular twins has an importance, among other things, in the study of the hereditary nature of certain illnesses. Zygosity is, however, not at all trivial after a time shortly following the birth. Twins are usually supposed to be monozygotic, and various examinations (concerning anthropometry, blood-group, trace elements, etc.) either state dizygosity or do not alter the original supposition of monozygosity. However, the latter does not exclude dizygosity. For making the diagnosis more accurate and convenient, Forrai suggested to study the voice of twins. For this purpose records, of identical texts lasting for appr. 1 minute each, by 117 twin-pairs were supplied by Forrai and Lubi (1973). The essence of the approach based upon speech recognition and speaker identification was summed up in [9].

It was shown in numerical terms, that by aptly defining a "distance", certain acoustic characteristics of the speech of two persons are, in the average, closer if the two speakers are twins.

The distance between the individuals i and j along a characteristics J is defined as

$$D_{ij}(J)=\frac{J_i-J_j}{J_i+J_j}.$$

The results of the investigations are as follows:

Characteristics investigated, J_i**	Average "distance", $E\{D_{ij}(J)\}$***	
	for the set of twins	for the set of non-twins
Explosive duration averaged over the individuals	0.12	0.26
Formant frequency	0.05*	0.07*
Mean fundamental frequency	0.04	0.06

* Averaged also over the set of all three formats of the Hungarian vowels á and é (like "i" in shine and "a" in able).

** Basic data were supplied by the system based on the PDP 11/40.

*** E denotes expected value.

From these data the acoustic relationship of twins is obvious.

Chances for a successful distinction between mono- and dizygotes can already be assessed on the basis of the simple difference in the average fundamental frequencies measured on the twins. The differences measured on 49 male twin pairs are shown in Table 1.

From the table it becomes clear that there is a chance for distinction, but a single characteristic is not enough for this purpose.

Table 1

Diff. in. average fund. freq.	Non-acoustic classification	Diff. in. average fund. freq.	Non-acoustic classification
0.8	M	7.9	M
1.1	M	8	D
	D	8.2	D
1.2	M	8.3	M
1.5	M	8.6	D
1.6	M	11	D
1.9	M	11.9	D
	M	12	D
2	M		M
2.1	M	14.5	D
	M	16.6	D
3.1	M	18	M
3.2	D	21.6	D
3.4	M	23.4	D
4.1	M	23.7	M
	M	23.8	D
4.5	M	27.9	M
4.6	M	30.5	D
5.1	M	33.5	M
5.7	M	33.7	M
	D	38.9	D
5.9	D	39.1	D
6	D	52.3	M
6.5	M	59	M
6.7	D		

(Basic data were supplied by the microprocessor system.)

Related to this, two promising examinations are in preparation for publications in the near future.

— The first one uses a multivariable discrimination analysis based on the average duration of the explosive "*k*" at the beginning of a word as well as the evaluation of the monochorus.

— The other one classifies vectors, with nine-components, using nearest neighbour decision. The weightings of the components are optimized by a learning algorithm. The nine components include three formants for "á", the same for "é", expected value and standard deviation of the pitch and the average of explosive durations.

5. Acknowledgement

The majority of the results described has been obtained in collaboration with university students in the frame of diploma work, scientific students' circles or independent laboratory work, lasting generally several months for every student. The students (in the meantime graduated as engineers) contributing most valuably are listed here in the sequence of their graduation: Ágnes Füredi, Péter Nagy, György Takács, Erika Garazsi, János Gönci, Géza Molnár, László Rumi, Rudolf Pannuska, Gábor Schmidt, László Dankó and Zoltán Kormány. A special thank is due to Dr. Rudolf Földváry, János Gönci, Dr. László Osváth and especially György Takács for their many good ideas.

The conditions for the students' work and thus the chance for the birth of the feature extraction system had been created on the PDP-8 by Dr. Lajos Pásztorniczky and Tamás Hetényi, on the PDP 11/40 by Dr. László Schnell and his coworkers, while on the autonomous microprocessor system by Dr. Sándor Csibi.

Summary

An account is given of speech processing, and of some new aspects on its present state. In addition, an acoustic feature extraction system based on a combined computer-microprocessor assembly is described.

Research results are publisheu on the determination of twin-zygosity (whether twins are monovular or binovular) by means of speech processing, focusing on the results of the classification based on comparing pitches and durations of explosives. The immediate aim of these investigations is to help in establishing the inheritability of certain deseases.

References

1. Helmholtz, H.: Die Lehre von den Tonempfindungen, Braunschweig, 1913.
2. Békési, Gy.: Experiments in Hearing, McGraw Hill. 1960
3. Tarnóczy, T.—Radnai, J.: Eine Möglichkeit automatischer Erkennung von Vokalen, Proc. VII. ICA, Vol. III., Budapest, 1971, p. 61—64
4. Földváry, R.—Gordos, G.: A new hypothetic model for pitch recognition in human speech, (In Hungarian), Híradástechnika, Vol. XXV. (1974), No. 11, p. 344
5. Gábor D.: Acoustical Quanta and the Theory of Hearing, Nature, Vol. 169, (1947) p. 591
6. Harkevich, A. A.: Spectra and analysis, Consultants Bureau, New York, 1960
7. Grobben, L. M.: Appreciation of short tones, VII. ICA, Budapest, 1971, Vol. 3 p. 329
8. Tarnóczy, T.—Vicsy K.: Some Remarks on the Perception of Voiceless Stopconsonants, Acoustica, Vol. 43, 1979, No. 2, p. 167
9. Forrai, Gy.—Gordos, G.—Lubi, B.: Preliminary Report on Voice-Based Discrimination between Monozygotic and Dizygotic Twins, Proc. Phys. Inst. ELTE (in press)

dr. Géza Gordos, H-1521 Budapest