JOURNAL ON

COMMUNICATIONS

VOLUME XLIII. JULY-SEPTEMBER 1992

SPEECH PROCESSING

JOURNAL ON COMMUNICATIONS

A PUBLICATION OF THE SCIENTIFIC SOCIETY FOR TELECOMMUNICATIONS, HUNGARY

Editorial office
Gábor Áron u. 65.
Budapest, P.O.Box 15.
Hungary, H-1525
Phone: (361) 135-1097
(361) 201-7471
Fax: (361) 135-5560

Subscription rates
Hungarian subscribers
1 year, 12 issues 2900 HUF, single copies 360 HUF
Individual members of Sci. Soc. for Telecomm.
1 year, 12 issues 480 HUF, single copies 60 HUF
Foreign subscribers
12 issues 90 USD, 6 English issues 60 USD, single copies 15 USD
Transfer should be made to the Hungarian Foreign Trade Bank, Budapest, H-1821, A/C No. MKKB 203-21411

JOURNAL ON COMMUNICATIONS is published monthly, alternately in English and Hungarian by TYPOTEX Ltd. H-1015 Bp. Batthyány u. 14., phone: (361) 202-1365, fax: (361) 115-4212. Publisher: Zsuzsa Votisky. Type-setting by TYPOTEX Ltd. Printed by HUNGAPRINT, Budapest, Hungary

HUISSN 0866-5583

EDITORIAL

If we accept the next paragraph as the definition of speech processing than speech processing looks back on a history of at least 201 years in Hungary because Farkas (von) Kempelen published his work on his speaking machine in 1791 [1].

Speech processing/technology aims at substituting or realizing by artificial means, one or more members of the natural speech chain consisting of the taking human, the air as transmission medium and the listening human.

If speech processing focuses only on the two human ends of the natural speech chain than it is realizing "artificial speech functions". Farkas von Kempelen did exactly this by devising an instrument that produced speech if the controls of the instrument were manipulated skilfully.

György Békésy was interested in the other end of the speech chain [2]. By his experimental investigations in Budapest he significantly improved our understanding of the operating principle of the inner ear and its role in the hearing process.

Dennis Gábor, a born Hungarian made an interesting effort in the 40's to comprehend the discrete nature of human speech [3].

The combination of filters and detectors capable of distinguishing between a fair number of speech sounds indicated in Tamás Tarnóczy's activity [4] the dawn of modern electronic speech processing in the early 50's. T. Tarnóczy's broad-spectrum-knowledge in acoustics has been and is a tremendous support for the speech processing community in Hungary.

From time to time the emphasis in speech processing is shifted to the middle element of the speech chain, i.e. to the transmission medium. This was indicated in Hungary by the development of a generator producing a random signal that simulated speech in terms of both the power spectral density and the first order (amplitude) distribution [5].

The turn of the decade from the 60's to the 70's was dominated by debates of "philosophycal" nature. While some people thought that the only way to artificial speech recognition was the imitation of the human hearing mechanism others argued that the message was encoded in the speaking organs, too, and if features of the speaking organs could be extracted from the speech waveform, as they indeed could be e.g. by linear prediction, a step towards speech/speaker recognition could be made ("speech recognition by the analysis of speech production"). Pessimism was supported, among other, by the realization that human ear performed pitch determination much better than any linear device could do at all. Optimism was strenghened, among others, by showing that digital processing might be able to perform that job equally well [6].

Formal organization of speech processing activities, and emerging of scientific schools in this field started in the late 60's and gained momentum in the 70's. First, probably, was the Speech Research Laboratory at the Dept. of Telecommunications and Telematics (and its predecesor). Technical University of Budapest (SRL/DTT/TUB). With long and rich traditions in linguistics, speech analysis and telephonometry, the Phonetics Department of the Linguistics Institute, Hungarian Academy of Sciences (PL/LI/HAS), the "Békésy György" Acoustics Reserach Laboratory, Hungarian Academy of Sciences (ARL/HAS) and the Research Institute of the Hungarian Telecommunications Company (RI/HTC), resp., became active in speech processing. These four groups, while maintaining their individual and historically founded characteristics, gradually developed a strong cooperation. Graduates of DTT/TUB were clearly contributing to these partnerships. A group in the KFKI Institute for Measurement and Computing Technics, Hungarian Academy of Sciences (KFKI/HAS), while partly relying on cooperation with PL/LI in the basics of text-to-speech conversion, emerged on the scene with special emphasis and expertise in speech synthesis for the blind. The Telecommunications Research Institute (TKI) appeared soon with its strength in low-bit-rate coding.

Even if the bulk of research and development in speech processing has been done in the laboratories mentioned above there are a considerable number of other contributing places and individuals, too.

And, of course, there are a few big manufacturing companies and a dozen small businesses offering speech technology products on the market, mostly based on the work of R & D places mentioned earlier.

From the many branches of speech processing it is probably speech synthesis that has attracted the biggest research effort and has led to the most applications. On the R & D field multilingual text-to-speech (TTS) converter (DDT-LI), a telephone number announciator (RI), a public announcement system (DTT/TUB), a TTS system specially designed for the blind (KFKI) and another that uses only a personal computer with a D/A (DTT), have to be mentioned by all means. On the market at least five manufacturing companies (one foreign, applying the patent of DTT-LI) indicate the state-of-art in Hungarian speech synthesis.

DTT and ALR developed isolated word speaker dependent recognition systems independently in the early 80's, than joined forces and have been producing successful prototypes of speaker independent and connected word versions. By cooperation of RI and LI using neural networks and speech data base for testing purposes, continuous speech recognition became also a target.

RI and DTT are active in the assessment of the quality of encoded—decoded speech.

Help to the handicapped is the motivation for a speech visualizer developed at DTT.

Space in this issue does not permit the presentation of all the applications of speech technology developed in Hungary. Still, without mentioning at least some of them the picture would be severely deficient. LI developed a synthesizer based audio-tester that is extensively used especially for children. DTT developed a speech detector that indicates the presence of the speech in severely noisy environment and also a speech enhancement device. Both devices are manufactured and are rather widely used.

DTT devised and implemented several speaker recognition algorithms and is working on finding their best combination. TKI's and DTT's results in low bit rate coding

seem to deserve attention. The guest editor's own favorite is to determine from an approximately one minute long speech recording, corresponding to a pair of twins whether they are fraternal or identical twins [7].

Even if the objective of speech processing is the practical realization (!) of one or more members of the natural speech chain by artificial means pure scientific results are also products of speech technology. Again, this issue is far too limited even to indicate all of them. The guest editor has had great difficulty in selecting only one/two purely theoretical papers from the many that are worth for publication. Under these circumstances he takes, with the readers' kind permission, the liberty to mention only his introduction of the notions of the time-warped-average/clustering and the characterization of speech by LPC-pole-trajectories and their resida [8].

Researchers feel that they always lack funding. This is very much the case also in Hungary. Still, credit is due to the National Committee for Technological Development (OMFB) and to the national Scientific Research Fund (OTKA), as well as to the Hungarian Telecommunications Company (MATÁV), BEAG, Telefongyár and others for providing funds, for the projects in speech technology.

No scientific and/or professional community can progress without communicating with each-other and with the rest of the world. At home, the Hungarian community of speech processing takes advantage of the Conference on Acoustics (separate speech processing sections were held at the 7th in 1982, the 8th in 1985, the 9th in 1988 and the 10th in 1991), LI organized the Conference on Speech Research '89 in 1989. These meetings have always attracted large international attendance. The Speech Technology Forums, organized yearly, are aimed at informing each other within the country.

Professional associations like the Scientific Society for Telecommunications (HTE), the Society for Optics, Acoustics and Precision Mechanics (OPAKFI) and the John von Neumann Computing Society (NJSZT) regularly organize meetings on speech technology. These three societies established the Kempelen Farkas award for outstanding contributions to speech processing. Laurates were T. Tarnóczy, G. Gordos in 1991 and Gy. Takács, G. Olaszy in 1992.

Of course, members of the community regularly publish papers in highly reputable international journals and conferences.

Across the borders span international cooperations. The Technical University of Vienna, the Research Institute of Seibersdorf, Austria, the German Blindenstudienveranstalt, the Intitut de la Communication Parlée, Grenoble and KTH Stockholm are just a few examples of long standing partnerships.

Without the constant fertilizing influence of these partnerships it would have been much more difficult for me, as guest editor, to have a really rich choice and to fill this issue with papers with extreme ease. I would like to thank all the authors for their papers and the cooperation also in reducing the sizes of their manuscripts. I have to apologize to those authors who sent their manuscripts but did not get published in this issue and to those potential authors who would really deserve a place in this issue. Finally, I would like to thank the Editor in Chief and the Editorial Board of the Journal on Communications for the great honour of inviting me as guest editor of this issue.

GÉZA GORDOS

REFERENCES

[1] Kempelem, W. v., "Le Mechanisme de la Parole, suivi de la Description d'une Machine Parlante", J.V. Degen. Vienna, 1791.

[2] Békésy, Gy. v., "Über die Schwingungen der Scheckentrennwand beim Präparat and Ohrmodell", Akust. Z. Vol. 7. 1942. pp.173–186.

[3] Gabor, D., "Acoustical Quanta and the Theory of Hearing", Nature, Vol. 169. 3rd May 1497, pp.591–602.

[4] Tarnóczy, T., Radnai, J., "Eine Möglichkeit Automatischer Erkennung von Vokalen", Proceedings of the 7th Int. Congr. on Acoustics, Vol. 3. pp.61–64. Akadémiai Kiadó, Budapest, 1971.

[5] Tarnay, K., Gordos, G., Melegh, J., "Load Simulating Noise Generator for Testing Carrier Telephone Systems", Budavox Telecommunication Review, No. 3/4, 1966. pp.12–19.

[6] Földváry, R., Gordos, G., "Hypothetical Model of Human Pitch Perception", Híradástechnika, Vol. XXV. No. 11. 1974Nov., pp.344–348 (in Hungarian)

[7] Forrai, G., Gordos, G., "A new acoustic method for the discrimination of monozygotic and dizygotic twins", Acta Paediatrica Hungarica, Vol. 24, No. 4., 1983. pp.315–321.

[8] Gordos, G., "New Feature Extraction Methods and the concept of Time-Warped Distance in Speech Processing", Proc. of Globecom'91, Phoenix, 1991. pp.21.7.1–21.7.5.

Géza Gordos received his M.Sc. in Electrical Eng. and Ph.D. in Telecommunications Eng. from the Technical University of Budapest (TUB) in 1960 and 1966, resp., and his C.Sc. from the Hung. Acad. of Sc. in 1978. Since 1976 he has been the Head of Dept. of Telecom. and Telematics (and its predecessor), and since 1972 the founding Head of Speech Research Laboratory, both at TUB. He is also serving as the Vice Rector for this University and as the President of the Hungarian Scientific Society for Telecommunications. He is founding member of both the Hungarian Academy of Engineering and the Telecommunications Systems Committee of the Hungarian Academy of Sciences. Earlier he worked, among others, for the Research Institute of the Hungarian Telecom. Co.; the University of Salford, U.K.; the UNESCO; the Imperial College, U.K.; the Telecommunication Work Co. His research and practical work has been focusing on speech and data transmission, telecom. systems and speech technology.

CONTINUOUS SPEECH RECOGNITION ON ACOUSTIC-PHONETIC LEVEL

K. VICSI
ACOUSTIC RESEARCH LABORATORY
OF THE HUNGARIAN ACADEMY OF SCIENCES

P. BERÉNYI
ACOUSTIC RESEARCH LABORATORY
OF THE HUNGARIAN ACADEMY OF SCIENCES

J. PARUCH
DEPT. OF TELECOMMUNICATIONS
AND TELEMATICS, TUB

G. LUGOSI
DEPT. OF MATHEMATICS, FACULTY OF
ELECTRICAL ENGINEERING, TUB

In Hungary, research on continuous speech recognition has started three years ago in cooperation with Acoustic Research laboratory of HAS and Department of Telecommunications and Telematics, Technical University of Budapest. In this paper the methods developed and the results achieved are presented. The speech signal was preprocessed according to a simplified auditory model. For further processing two different technics were used. One is a rule based on segmentation and recognition, the other is based on a statistical method, the so-called hidden Markov model. These two approaches are discussed separately in detail. In a realization of a connected word recognizer a combination of these two methods are presented.

1. INTRODUCTION

Over the last two decades, people in research have worked on development of large vocabulary speech recognition systems, used as voice activated typewriters. Now, by the 90s, such systems have been developed in the leading speeh laboratories. These systems are speaker adaptive, their dictionary has some thousand words, however, words should be separated by short intervals [1].

Computational requirement for speech processing is very high, but technology tries to follow the requirements. The increasing capability and the increasing density of single chip digital signal processors are shown in Fig. 1 as a function of time [2].

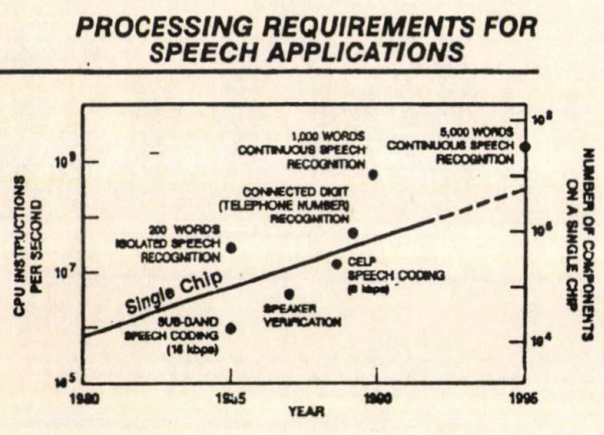


Fig. 1. Computational requirements for selected speech processing algorithms (Points). Processing capability and component density for single-chip digital signal processors as a function of time (Line). J.L. Flanagan, 1991.

There are language-independent and language-specific aspects in speech research. Europe is a mosaic of small countries, each one having one or several different languages. Speech recognizers for English cannot help them very much. Automatic language and voice processing may be considered as a priority within the framework of building a community of countries, without destroying the cultural heritage and future of each of them [3]. Countries in Western Europe have seen the importance of development of speeh technology and made big efforts at the national and international level. Several Western European transnational research and development projects and networks exist, most of them are financed by the EC. New and really good results show that these cooperations between the countries, and the appropriate financial support has been fruitful.

Till now, Hungary has been excluded from this international cooperation, but the scientific and technical manager group of Hungary has no doubt about the importance of speech research and tries to give financial assistance ("National Scientific Research Find", "Natioal Technical Development" grants, etc.) according to the moderate economical possibilities of the country. Some years ago speaker dependent isolated word recognizers for some hundred words were developed in Hungary [4], [5], [6], [7], and now they are on the market. Speaker independent systems are under development. Moreover, three years ago research on continuous speech recognition had been started in cooperation with the Acoustic Research Laboratory, HAS and the Department of Telecommunications and Telematics, Technical University of Budapest. This year the Phonetic Laboratory HAS joined the team.

2. CONTINUOUS SPEECH RECOGNITION

In order to illustrate the difficulty of automatic speech recognition (ASR), we emphasize only two main problems: segmentation and adaptation. The first problem is connected with the fact that the action of articulatory mechanism in continuous speech is a continuous flow of phonetic information where there are no obvious physical indications of the boundaries of words, syllables or other phonetic units. Adaptation means that the computer must be adapted to the different voices used as input. It is a very difficult problem to design a recognition algorithm that can cope with the myriad ways different speakers pronounce the same sentence or, indeed, to interpret the variations that a single speaker uses pronouncing the same sentence at different occasions.

In continuous speech recognition, the acoustic properties of a given word can change significantly, depending on its position in a sentence. The acoustic properties of a word can also be modified by the adjacent words. Finally, syntax, semantics and knowledge of the matter spoken of can also directly modify the speech signal.

Our continuous speech recognition system has been developed, in the first phase, to recognize speech on phonetic level, and construct a database for the Hungarian language. The starting algorithm is shown in Fig. 2. A simplified auditory model is used for acoustical analysis and further processing is done by an IBM 386 or MICROVAX Q.II. This further processing of the signal is performed in two different ways. One method is the so-called rule based segmentation and recognition, the other is based on a statistical method, the so-called hidden Markov model. These two approaches are normally discussed separately, but they are not too different. In a rule based system, researchers use rules based on extremely flexible "statistical" analysis made by the human brain. In a realization of a connected word recognizer, a combination of these two methods are presented.

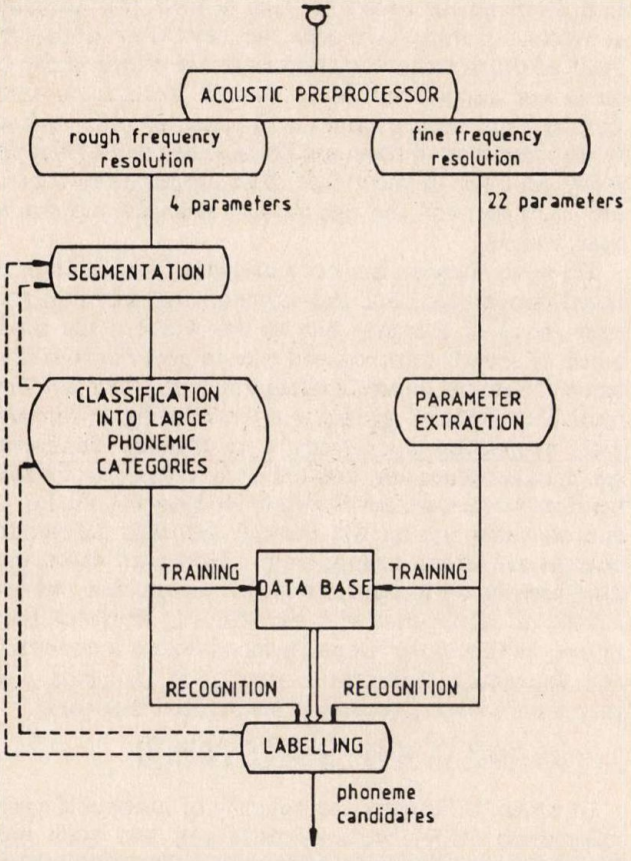


Fig. 2. Continuous speech recogniotion on acoustic–phonetic level

2.1. Auditory model for acoustic analysis

Our acoustical preprocessor is based on an auditory model [8]. This preprocessor has been used for isolated word recognition with success. The auditory model is based on the results of psychoacoustic researchers [9], [10], [11]. The model analyzes the speech signal with approximately the time, frequency and intensity resolution of the human peripheral auditory system during speech perception. The data are valid for average speaking rates and average speech intensities between 100 and 8000 Hz. The frequency resolution is 1 Bark. We use critical band filtering. The filters have asymptotic slopes reflecting the filter characteristics of the human ear (Fig. 3). The time resolution is 10 ms which means that in every 10 ms, 20 filter outputs, the total energy, the energy below 1 kHz and the zero crossing counter outputs are sampled. (The last 3 parameters do not belong to the auditory model. These are used when quick decision is necessary.) Fig. 4 shows the change of the total energy, energy below kHz, the zero crossing number with time and the filter outputs as a function of time. In the Figure, the following phonetically blanced Hungarian sentence fragment is presented: "A falatozóban sör, bor, üdítőital". This sentence fragment is shown on all the figures, so that the reader can follow the different processes in the pictures.

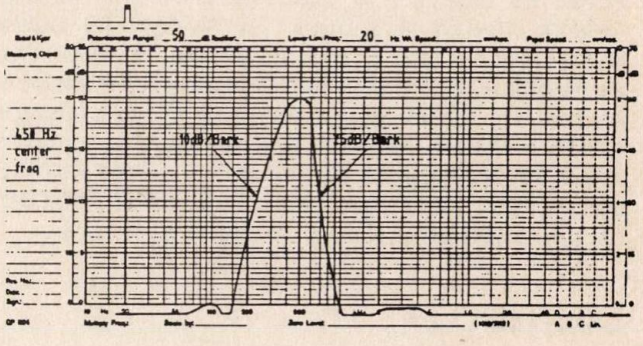


Fig. 3. Filter characteristics

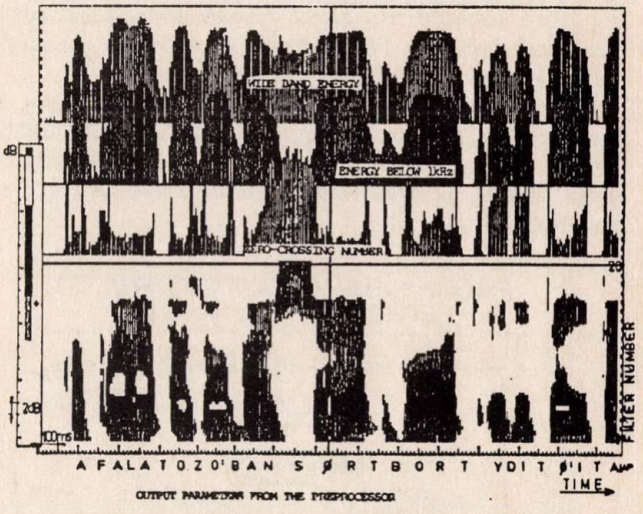


Fig. 4. Output of the acoustical preprocessor

2.2. Rule based segmentation and recognition

In a rule based system, the knowledge of experts formulated in rules are used for the segmentation, training and classification. In our recognition system (Fig. 2), rough frequency resolution (energy of 4 special bandfilters) is used for automatic segmentation and for the classification of frames into large phonemic categories. While rough frequency resolution is absolutely suitable for segmenta-

tion, fine frequency resolution, (all filter outputs of auditory model) is necessary for labeling.

2.2.1. Segmentation

The success of recognition results depends considerably on the success of segmentation. To obtain a good result, not only the segmentation technique is important but also the choice of suitable phonemic units. The language, the aim of recognition, the preprocessing system and many other things determine which phonemic units are the best.

Here the phonemoid was chosen as a unit of segmentation. These segments are generally phonemes but are adapted to the acoustic features obtained by the auditory model, sometimes the segment is shorter, sometimes longer than a phoneme. For the rough frequency resolution the following filters were used:

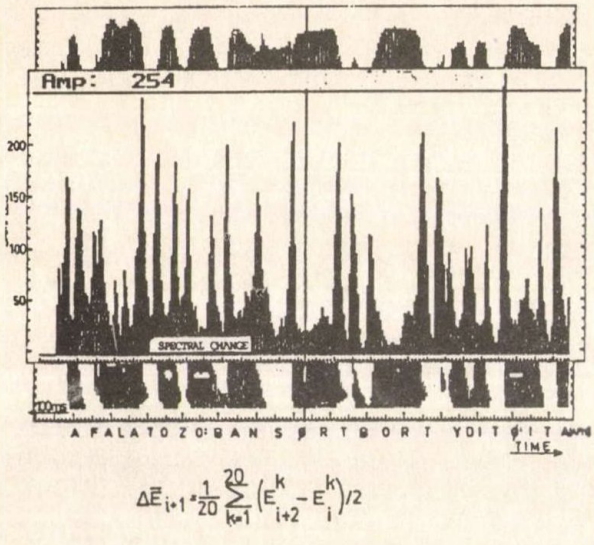
resonant filter:	100 – 1000 Hz,
spirant noise filter:	1000 – 7700 Hz,
nasal filter:	100 – 800 Hz,
pitch filter:	100 – 400 Hz,.

These are not implemented in hardware but are obtained as the sum of some filter outputs of the model.

The segmentation has been done in more steps, and the mistakes at a lower stage may be corrected on the base of the results obtained at higher stages. At the first stage of segmentation, a simple spectral change measure is used:

$$E_{i+1} = \frac{1}{4}\sum_{j=1}^{4} |(E_{i+1}^{j} - E_{i}^{j})|/2,$$

where E_i^j represents the log energy of the j-th "rough" filter output at the i-th frame. The spectral change is shown in Fig. 5 in case of our reference sentence.



$$\Delta\vec{E}_{i+1} = \frac{1}{20}\sum_{k=1}^{20}\left(E_{i+2}^{k} - E_{i}^{k}\right)/2$$

Fig. 5. Spectral change in time (i: frame numbers; k: filter numbers)

At the second stage, segments are classified into larger phonemic categories, as silence, aspirant noise, burst, vowels, nasals, laterals, etc.

At the following stage, phonemic and phonological rules are taken into consideration and segment boundaries may be corrected. For example: silence, or closure periods of stops were decided when the total energy was below a moving threshold, as shown in Fig. 6. The problem is whether these segments are really silence or they belong to stops or affricates. This can be decided if the following segment is examined. In those frames where the total energy is greater than the energy below 1 kHz, burst and/or aspirant noise must be present, and the silence before these segments must be part of a stop or affricate.

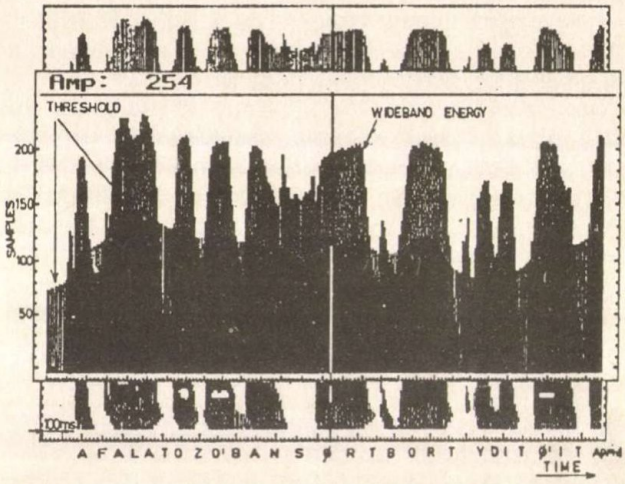


Fig. 6. Moving threshold for determining the silence periods

Taking into consideration several rules similar to the one mentioned above, the automatic segmentation was performed, and the results were compared with hand-made ones. One example is shown in Fig. 7. This segmentation method works well in nearly all utterances, except laterals and voiced stop-plosives which are frequently drawn together with vowels [12]. The described method was tried on other European languages, e.g. German, Finnish and Swedish. Similar results were obtained in these cases too. This is not surprising, since only general phonetic rules have been used till now.

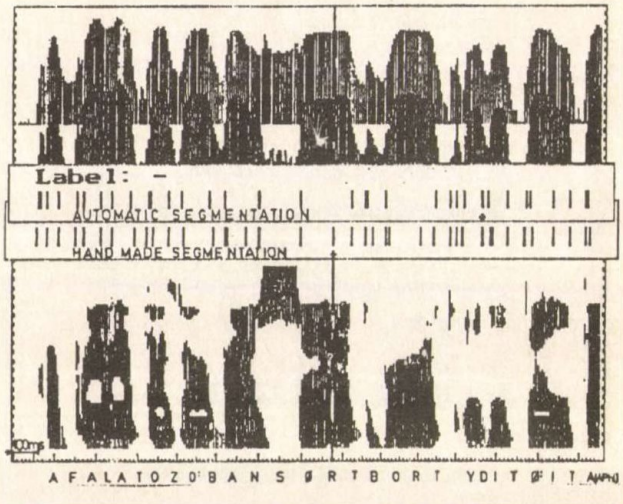


Fig. 7. Rule based segmentation compared with one done by hand. The vertical axis is similar to that of Fig. 4.

2.2.2. Training, automatic labelling for training

Every speech recognizer needs some training procedure: in statistical computations one needs a lot of data to build up good models of the units, in knowledge based systems a large data base must be built up, while in the neural network approach one needs a lot of well segmented and labelled examples to train the net. Generally hand made segmentation and labelling techniques are used for training but they are time consuming. In the Acoustic Research Laboratory, an automatic training procedure was developed for our system and for the Hungarian language, but finally it was realized that the technique is good for other European languages too.

Our philosophy was the following: If you know the labels of the trained sentences, simultaneously with the automatic fonemoid based segmentation, the labelling of each single segment can be made. Moreover, mistakes in the segmentation may be corrected. This technique makes the job of an expert easier.

If the automatic segmentation was correct the labelling of the segments were made without any problem. But in practice, the automatic segmentation has mistakes, and the labelling could be quite wrong.

Our method is described in the upper part of Fig. 8. The system makes the best decision in the case of the following categories: aspirant – noise (SPIRR), vowel (VOW), nasals – laterals (NALI) and silence (SIL) (when 4 phonemic categories are used). According to our preprocessing system, the 4 categories can be represented in a two-dimensional field by the following vectors:

PHONEMIC CATEGORIES:	(SPIR)	(VOW)	(NALI)	(SIL)
PARAMETERS:	4,8	8,10	5,8	0,0

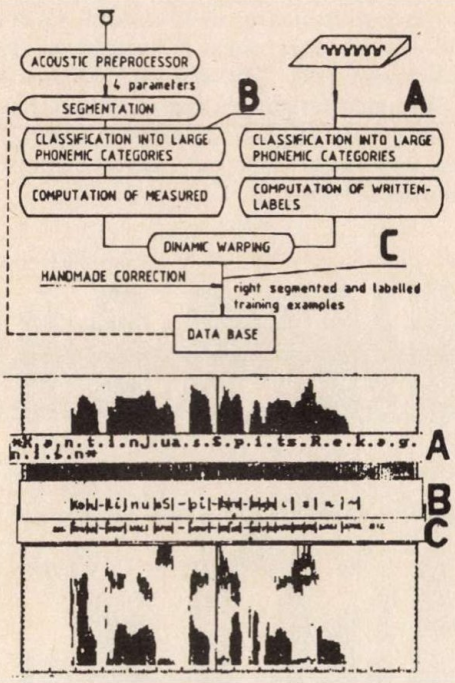


Fig. 8. Automatic training procedure

Categorization of the automatically segmented units:

In the case of the automatic segmentation and classification, all frames inside the segments get these numbers according to the phonemic category the frame belongs to, so the whole segment is charaterized by the following two-dimensional vector:

$$\text{AUTomaticPHoneme}[j, l] = \sum_{n=1}^{\substack{\text{frame number} \\ \text{of the segment}[j]}} \text{PARAMETER}\,[\text{RawLabel}[j, n], 1]$$

where j is the ordinal number of segment in the utterance, n is the ordinal number of the frame within the segment, RawLabel$[j, n]$ is the code $(1 \ldots 4)$ of the phonemic category the n-th frame of the j-th segment belongs to, and PARAMETER$[k, 1]$ $(k = 1 \ldots 4, l = 1, 2)$ is the l-th coordinate of the two-dimensional vector assigned to the k-th phonemic category. We can see that this parametric assignment is additive in the following sense: the vector belonging to a segment is the sum of vectors belonging to frames in that segment. So if the boundary between two segments is missed accidentally, the vector assigned to this composite segment is the sum of the two vectors which would belong to each segment if the automatic segmentation process could find the boundary between them.

Categorization of the written labels:

Ranging probabilities in the 4 phonetic categories of all phoneme-like units can be based on an earlier statistical examination [12] (10 phonetically balanced sentences, 464 phonemes). However, an expert who knows the system quite well can give these probabilities without any detailed computation.

For example, the Hungarian phonemes are the following:

PHONEMES:

('A', 'AA', 'B', 'C', 'CS', 'D', 'DZS', 'E', 'EE', 'F', 'G', 'GY', 'H', 'I', 'J', 'K', 'L', 'M', 'N', 'NG', 'NV', 'NY', 'O', 'OE', 'P', 'R', 'S', 'SZ', 'T', 'TY', 'U', 'UE', 'V', 'Z', 'ZS', 'PR', 'MB', 'ND', 'TR')

The phonetic transcription itself and the rules to transform it into a string of phonemoid units are evidently dependent on the language.

PHONEMEMOID:

('A', 'AA', 'B', '-B', 'C-', '-C', 'CS-', '-CS', 'D-', '-D', 'DZ-', '-DZ', 'DZS-', '-DZS', 'E', 'EE', 'F', 'G-', '-G', 'GY-', '-GY', 'H', 'I', 'J', 'K-', '-K', 'L', 'M', 'N', 'NG', 'NV', 'NY', 'O', 'OE', 'P-', '-P', 'R', 'S', 'SZ', 'T-', '-T', 'TY-', '-TY', 'U', 'UE', 'V', 'Z', 'ZS', '*', '#', 'PR-', '-PR', 'NG-', '-NG', 'MB', 'o-MB', 'ND', '-ND', 'TR-', '-TR')

(the # and * means coarticulation noises)

RANGING PROBABILITIES:

((0,0,0,6), (0,0,0,6), (3,0,3,0), (0,2,2,2), (5,1,0,0), (0,6,0,0), (5,1,0,0), (0,6,0,0), (3,0,3,0), (0,2,2,2), (3,0,3,0), (0,5,0,1), (3,0,3,0), (0,5,0,1), (0,0,0,6), (0,0,0,6), (3,3,0,0), (3,0,3,0), (0,2,2,2), (3,0,3,0), (0,5,0,1), (3,3,0,0), (0,0,1,5), (0,0,3,3), (5,1,0,0), (0,3,3,09), (0,0,3,3), (0,0,5,1), (0,0,5,1), (0,0,5,1), (0,0,5,1), (0,0,5,1), (0,0,0,6), (0,0,0,6), (5,1,0,0), (0,2,2,2), (0,0,3,3), (0,6,0,0), (0,6,0,0), (5,1,0,0,), (0,3,3,0), (5,0,1,0), (0,5,1,0), (0,0,3,3), (0,0,1,5), (2,2,2,0), (0,5,0,1), (0,5,0,1), (0,3,3,0), (0,5,1,0), (5,1,0,0), (0,1,3,2), (0,0,5,1), (0,2,2,2), (0,0,5,1), (0,2,2,2), (0,0,5,1), (0,2,2,2), (5,1,0,0), (0,1,3,2))

This matrix gives the conditional probability of a frame belonging to a segment to have a particular RawLabel if that segment is labelled by a given phonemoid unit. The numbers should be normalized by 6.

AVERAGE FRAME NUMBER:
array[1 . . . 60] of byte =
(11, 11, 9, 1, 9, 5, 9, 5, 9, 1, 9, 5, 9, 5, 11, 11, 9, 9, 1, 9, 5, 9, 11, 5, 9, 1, 5, 5, 5, 9, 5, 5, 11, 11, 9, 1, 2, 9, 9, 9, 1, 9, 5, 11, 11, 9, 9, 9, 4, 2, 9, 3, 9, 1, 9, 1, 9, 1, 9, 3)

This vector shows the average number of frames in a segment labelled by and appropriate phoneme-like unit. We tried to make the computation as simple as possible, so we use only mean values and ignore spread. Each written phoneme is characterized by a two-dimensional vector:

$$\text{WRITtenPHoneme}[j,1] = \text{AVERAGE FRAMENUMBER}[n] * \sum_{k=1}^{4} \text{RANGING PROBABILITIES}[n,k] * \text{PARAMETER}[k,1]$$

where j is the ordinal number of segment in the sentence, and n is the code of the phonemoid label of it. ($l = 1 \ldots 2$)

Dinamic warping:

Having parametrized the string of segments obtained by automatic segmentation from the spoken utterance (AUTomaticPHonemoids) by the above mentioned two-dimensional vectors and also the segments obtained by replacing the written phonemes (orthographic character string) belonging to this utterance with a string of phonemoids (WRITtenPHonemoids), thus obtaining two strings of vectors of the same kind, the warping of them is still necessary. To this end, we used a slightly modified version of a standard dynamic warping algorithm:

$$g[i,j] = \min \begin{cases} g[i-2,j-1] & + \text{dist (AUTPH}[j],\text{WRITPH}[i-1]+\text{WRITPH}[i]) \\ g[i-1,j-1] & + \text{weight} * \text{dist (AUTPH}[j],\text{WRITPH}[i]) \\ g[i-2,j-2] & + \text{dist (AUTPH}[j-1]+\text{AUTPH}[j],\text{WRITPH}[i]) \end{cases}$$

where AUTOPH[i] and WRITPH[j] are the two strings of vectors characterizing the i-th segment in utterance and j-th segment in written sentence, respectively. The function 'dist' is the euclidean distance of the vectors having been normalized each of them by the length of the appropriate segment. The modification of the dynamic warping algorithm consists of the summation inside the distance function rather than outside, as usally should have been, but it is correct because of the additive property of the parametrization.

Having gone backwards on the route where the distance was minimal, the final labelling of the segmented units and the correction of the segment-boundaries are made.

Results:

In a data base of 10 phonetically balanced Hungarian sentences, the segmentation error was 9%, but no bursts were included in that statistics and the labelling error was the same. We have tried our method on other European languages, e.g. German, Finnish, and Swedish. Similar results were obtained in these cases too. This is not surprising since general phonetic rules are used in our method. If the matrices of PHONEMES, PHONEME-LIKE UNITS and AVERAGE FRAME NUMBERS are made according to the rules of a given language, you should obtain similar results.

2.3. Recognition based on a hidden Markov model (HMM)

In a hidden Markov model, a sequence of symbols is supposed to be produced by different sequences of states, and in general the symbols are observed but the state sequence is unseen. Once the parameters are known, the probability of transition can be computed as the product of the transition probabilities, and the probability of a symbol sequence is the sum of the probabilities of the paths that can produce in [13]. Training algorithms are available to automatically compute the parameters, and classical algorithms can be used to perform a graph search to find the most likely path [14].

Generally models represent speech units like words, syllables, etc. The model is based on statistical examination of the adequate number of utterances.

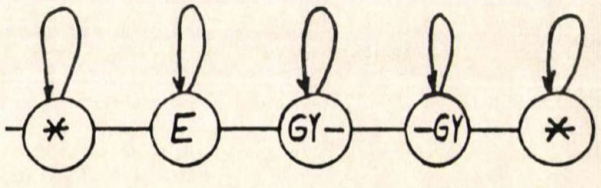


Fig. 9. The Markov model of the Hungarian word "EGY"

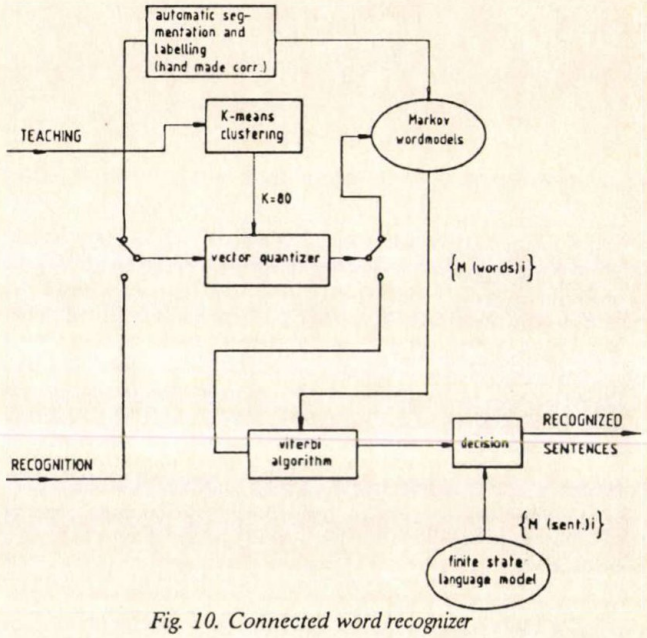


Fig. 10. Connected word recognizer

At the Department of Telecommunications and Telematics of the Technical University of Budapest, a special algorithm has been constructed. In this algorithm, as seen in Fig. 9, one model represents one word, and the states rep-

resent phonemoids (characterized in 2.2.1.) which build up the words. So that the number of states per word changes. This means that the number of the states is equal to the number of phonemoids in the word. To construct the models, segmented and labelled teaching material is necessary. This material was made by the technics described in 2.2.

The 23 entry vectors (obtained by the auditory model) were quantized by a vector quantizer designed by the K-means cluster technic [14]. After a detailed examination, the codebook of size 80 was found as an optimum. Various versions of the so-called Viterbi training algorithm were used to estimate the model parameters [15].

3. A CONNECTED WORD RECOGNIZER

For a recognition system designed for train or flight ticket reservation, a simple finite state language model was constructed. The number of words in the dictionary was 50, but the possible number of sentences is more than one million. During the training period, 150 segmented and labelled words were used to construct the models. The recognition system is shown in Fig. 10. Although the word recognition rate was fair (85%), the recognition on sentence level was close to perfect (100% for 14 sentences containing 58 words) due to the additional grammatical constraints.

REFERENCES

[1] Baker, J.M., "Large vocabulary speaker-adaptive continuous speech recognition research overview at Dragon Systems", *Proc. of EUROSPEECH 91.* Genova, 1991.

[2] Flanagan, J.L., "Speech Technology and Computing: A Unique Partnership", *Proc. of EUROSPEECH 91.* Genova, 1991.

[3] Marianio, J.J., "Speech Research and Technology: Advances in Europe", *Proc. of EUROSPEECH 91.* Genova, 1991.

[4] Vicsi, K., Lugosi, G. and Linder, T., "Search for Optimal Teaching Procedure and Warping Algorithms for Isolated Word Recognition Device", *Proc. of XIth IcPhS.* Tallin, 1987.

[5] Faragó, A., Gordos, G., Koutny, I., Magyar, G., Osváth, L. and Takács, Gy., "The VERBIDENT – SD2 isolated word machine recognizer" (in Hungrian), *Híradástechnika XXXIX. 1988/3.*

[6] Faragó, A., Gordos, G., Koutny, I., Magyar, G., Osváth, L. and Takács, Gy., "VERBIDENT – SD2 an isolated word recognizer", *Proc. of 9th Acoustic Conf.*, Budapest, 1989.

[7] Vicsi, K., "Are speech controlled machines viable?" (in Hungarian), *Magyar Elektronika*, VI.12. 1989.

[8] Vicsi, K., "Automatic Word Recognition System Using Some Results of Phychoacoustics", *Proc. 6. FASE Symp.* Sopron, 1986.

[9] Zwicker, E.: Psychoakustik. Springer Verlag, Berlin, 1982.

[10] Fourcin, A.J. et al., "Speech processing by man and machine, "Group Report" on recognition of complex acoustics signals, edited by T.H. Bullock. *Life Sciences Research Report of the Dahlem Workshop*, Berlin, 1977.

[11] Vicsi, K., "The Most Relevant Acoustic Microsegment and Its Duration Necessary for the Recognition of Unvoiced Stops", *Acoustica 48.* (1981) 53.

[12] Vicsi, K., Mattila, M. and Berényi, P., "Continuous Speech Recognition Using Different Methods", *Acoustica 71.* (1990), 152.

[13] Rabiner, L. R., Juang, B.H., "An Introduction to Hidden Markov Models", *IEEE ASSP Magazin*, (1986) 4.

[14] Jelinek, F., Bahl, L.R. and Mercer, R.L., "Continuous Speech Recognition: Statistical Methods", *Handbook of Statistics II.*, edited by H.P. Krishnaiad, North Holland, 1982.

[15] Forney, J.R., "The Viterbi Algorithm", *Proc. IEEE, 61.* (1973) 268.

Klara Vicsi graduated at the Faculty of Natural Sciences of Loránd Eötvös University of Sciences in 1971. She received the doctor's degree from Loránd Eötvös University of Sciences in 1982, and her Ph.D. from the Hungarian Academy of Sciences in 1991. Now she is the chief of the Speech Group at the Békésy Acoustic Research Laboratory of the Hungarian Academy of Sciences. From 1971 to 1982 she analyzed the acoustic parameters speech sounds, and examined the connection between the physical parameters and the psychological judgements. Since 1982 she is working in the field of speech recognition.

Péter Berényi studied mathematics at the Loránd Eötvös University of Sciences till 1980. He is working at the Békésy Acoustic Research Laboratory of the Hungarian Academy of Sciences since 1987 as a mathematician and computer programmer. He is interested in signal processing, computational theory of hearing and speech recognition.

Gábor Lugosi graduated in Electrical Engineering at the Technical University of Budapest in 1987, and received his Ph.D. from the Hungarian Academy of Sciences in 1991. Between 1987 and 1991 he worked at the Institute of Communication Electronics. He is currently an assistant professor at the Department of Mathematics at the Faculty of Electrical Engineering of the TUB. His main interests are mathematical statistics and information theory.

ACOUSTIC-PHONETIC RECOGNITION OF CONTINUOUS SPEECH BY ARTIFICAL NEURAL NETWORKS

GY. TAKÁCS

PKI TELECOMMUNICATIONS INSTITUTE
H-1097 BUDAPEST, ZOMBORI u. 2. HUNGARY

This paper describes an artificial neural network that recognizes phonemes in continuous speech, based on error back-propagation training. The recognition is performed by three connected nets. First, a coarse feature network is trained to recognize seven quasi-phonetic features from 10 ms spectral frames. The features need not be binary but may take any values between 0 and 1. The outputs of the feature net as well as the spectral outputs of the filter bank are used as inputs to the second net, the phone net, which recognizes phonemes. A seven frames wide symmetric window of the feature net output is used to include the context of the frame being classified. The outputs of the phone are also used as inputs to a segmentation network. Fifty sentences of one speaker were used for training the different nets, and ten more were used for testing. The features of the coarse net were recognized with 80% to 95% accuracy. Correct phones were recognized with 64% accuary and in 82% of the cases, the correct phone was among the best three candidates.

1. INTRODUCTION

This paper reports experiments concerning phoneme recognition in continuous speech. A recognition system has been developed and evaluated while the author was on a seven month visit as a guest researcher at the Department of Speech Communication and Music Acoustics at TU Stockholm.

Speech recognition based on phoneme-like units is a long-term research objective. It is attractive since it is inherently free from vocabulary limitations. This is of special importance in highly inflected languages, as Hungarian. Dozens of different forms of the same root word may occur, making speech recognition based on word size units encounter severe practical limitations.

The system reported in this paper mixes some well established techniques with some new, recently published elements from the area of neural networks to a novel combination in order to optimize a speaker dependent phoneme recognition system.

It is generally accepted that speech recognition cannot be solved purely on the acoustic–phonetic level. There is, however, no reason to transfer acoustically–phonetically solvable tasks to higher processing levels. A main objective of this project is to recognize phoneme-like elements on this bottom level as well as possible. The recognition system includes both automatic segmentation and automatic classification of speech.

The output of the acoustic–phonetic level of the system is a string of phoneme candidates. Each phonetic event is related to one or several candidates in parallel. The data format may be extended by including information on durations and probabilities of the phonetic candidates if necessary. The output has no rigid sequential structure so that simultaneous and overlapping events in the speech process can easily be described.

Usually the first step in the acoustic processing of speech is a spectral analysis using 10 to 20 ms time frames. A considerable part of single frames is very uncertain. A window having a duration of several speech frames slides across the preprocessed speech material in single frame steps.

The most traditional solution to windowing is using a section of a spectrogram-like representation [21]. This representation has a high redundancy. Since automatic speech recognition sooner or later seems to be limited by the amount of data or the speed of computation, data compression has a value in its own. In our system a new method has been introduced, partly for the purpose of data compression. It represents each 10 ms frame by some basic articulation related features that are calculated from the speech spectrum. These "coarse phonetic features" describe the manner and the place of articulation. Another purpose of this feature representation is to divide the phoneme classification task into two separately managed subtasks. In the final phoneme classification of a single speech frame, the values of these features within a 15 frame symmetrical window are used together with the spectral vector of the current central frame. Using this representation, the compression rate is approximately 5 when compared to the spectographic representation above.

The recognition of phoneme-like units is traditionally done in two sequential steps: segmentation and classification. However, an exact segmentation frequently needs information available only after the classification. Many of the papers reporting excellent recognition performance only deal with manually segmented speech samples. Thus a very problematic part of the processing is not included. Komori et al. have made experiments with automatic segmentation and they report 9.2% errors on segmentation [10]. We have tried to build a system that does not depend on manual segmentation during the recognition phase.

The system has a hierarchical structure, and the underlying idea is to separate the system into a language independent and a language dependent part. This will facilitate the adaptation to a new language. The phoneme recognition system is based on two hierarchically connected but separately trained neural networks. It is based on well functioning elements of the most frequently used neural net structures. We may to some extent rather leave it to the network to learn details than modelling explicitly. Still, we need a vast amount of general ideas, theoretical knowl-

edge and experimental results to be able to construct good network structures and training procedures. The structure of a neural network can be used as a carrier of our background knowledge.

The principles and basics of neural networks are not discussed in this paper. This information can be found in [18], [7], [8], [13], [14], [15], [16], [20].

2. FEATURES OF THE PROCESSED SPEECH MATERIALS

The first experiments were conducted using a Swedish speech database. Later, a Hungarian database was created and processed. Most tests were made by the recognition system adapted to a single speaker.

Sixty Swedish sentences constitute the first speech material. They are part of a speech material that has been used in several other studies and will be referred to as the INTRED-material [5], [17]. The material consists of sentences read in a natural way by a trained male speaker having a central Swedish dialect. the reading speed was rather fast, 13.1 phonemes/s. The speech signal is sampled at 16 kHz using a 6.3 kHz lowpass filter. The sentences are labelled by a human phonetic expert using both visual and auditory information [17].

A new speech material was created especially for these recognition experiments. A native Hungarian male speaker reads the text in a natural way. The average speed was 12.6 phonemes/s. This material was recorded and analyzed using the same technique as described for the INTRED-material, and will be referred to as the MAMO-material. In these experiments, the Hungarian phonemes were represented by a set of 49 elements.

3. THE BASIC STRUCTURE OF THE COMPLETE SYSTEM

In this Section, we will give an overview of the proposed system. Details regarding functions and designs will be discussed in Section 4. Continuously spoken sentences constitute the input to the system. The output of the acoustic–phonetic processing is a string of phoneme candidates and may be used as input to a language processing stage. The output of a complete system would be ordinary written text. The connections between the basic units and the different representations used are shown in Fig. 1.

The acoustic processing does not make use of any higher level information. All the processing related to syntax and semantics (e.g., lexicon, grammar) is included in the second box in Fig. 1. This paper only deals with the acoustic–phonetic processing in the first box. It consists of 8 basic units as depicted in Fig. 2.

The input to the system is the speech waveform, see number 1 in Fig. 3, where the horizontal axis of each representation has the same time scale, and the vertical lines indicate the manually marked phoneme boundaries. The smoothed output signal of the filter bank is connected to the inputs of the feature classification neural network, and the sampling interval is 10 ms. Compare representation 2 in Fig. 3, where each column describes one filter frame with 16 filter outputs. Low-frequency filters are at the bottom, and the size of each square is proportional to the filter output magnitude in dB.

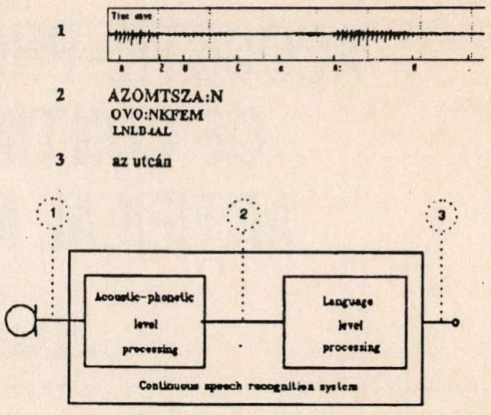


Fig. 1. The basic structure of the recognition system and the different representations used. The input is the speech waveform (representation 1). The output of the acoustic–phonetic unit is a string of phoneme candidates (representation 2). For each segment, the probability of the phoneme candidates is indicated by the character size, large for the first candidates and smaller for the second and third candidates. This level does not indicate any word boundaries. The output of the language unit is ordinary written text. The language procesing unit is not included in this study.

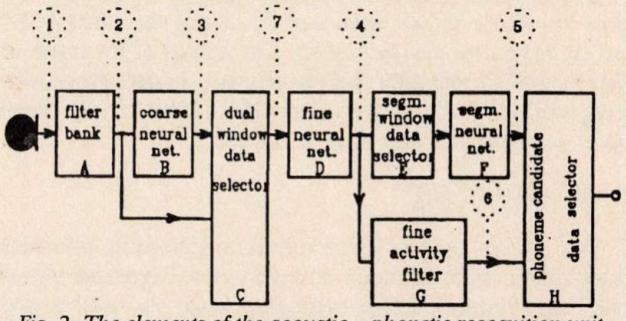


Fig. 2. The elements of the acoustic–phonetic recognition unit. Representations at different points of the system are shown in Fig. 3 and Fig. 4.

The output of the feature net is a seven element vector describing manner and place of articulation. This output is shown by representation 3 in Fig. 3 where the sizes of the squares are proportiolnal to the output activity of the feature net. The order of the features is the same as in Table 1 with the uppermost feature corresponding to feature number 1, voiceness.

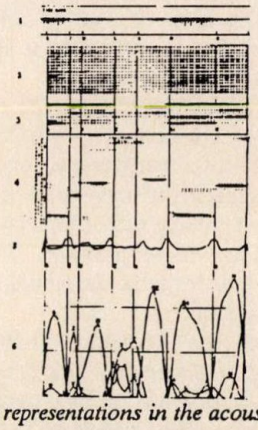


Fig. 3. Different representations in the acoustic–phonetic recognition system

Table 1. Coarse phonetic features for the Swedish phoneme set

Swedish technical alphabet	*IPA*	*voice-ness*	*noise-ness*	*nasal-ness*	*front-ness*	*central-ness*	*backness*	*vowel-ness*
.	(pause)	-	-	-	-	-	-	-
P	p (occ.)	-	-	-	-	-	-	-
T	t (occ.)	-	-	-	-	-	-	-
K	k (occ.)	-	-	-	-	-	-	-
p	p (burst)	-	+	-	+	-	-	-
t	t (burst	-	+	-	-	+	-	-
k	k (burst)	-	+	-	-	-	+	-
B	b	+	-	-	+	-	-	-
D	d	+	-	-	-	+	-	-
G	g	+	-	-	-	-	+	-
M	m	+	-	+	-	-	-	-
N	n	+	-	+	-	-	-	-
NG	ŋ	+	-	+	-	-	-	-
R	r	+	-	-	-	+	-	-
L	l	+	-	-	-	+	-	-
V	v	+	+	-	+	-	-	-
J	j	+	-	-	-	+	-	-
F	f	-	+	-	+	-	-	-
S	s	-	+	-	+	-	-	-
SJ	ʃ	-	+	-	-	+	-	-
TJ	ç	-	+	-	-	-	+	-
H	h	-	+	-	-	-	+	-
I	ɪ	+	-	-	+	-	-	+
E	e	+	-	-	-	+	-	+
Ä	ɛ	+	-	-	+	-	-	+
Y	ʏ	+	-	-	+	-	-	+
Ö	ø	+	-	-	-	+	-	+
U	ɵ	+	-	-	-	+	-	+
O	ʊ	+	-	-	-	-	+	+
Å	ɔ	+	-	-	-	-	+	+
A	ɑ	+	-	-	-	+	-	+
I:	i:	+	-	-	+	-	-	+
E:	e:	+	-	-	+	-	-	+
Ä:	ɛ:	+	-	-	+	-	-	+
Y:	y:	+	-	-	+	-	-	+
Ö:	ø:	+	-	-	+	-	-	+
U:	ʉ:	+	-	-	+	-	-	+
O:	ʊ:	+	-	-	-	-	+	+
Å:	ɔ:	+	-	-	-	-	+	+
A:	ɑ:	+	-	-	-	-	+	+

The more detailed phone classification neural net has a dual input window which includes the spectral vector describing the speech context by seven frames centred at the spectral frame, see Fig. 4. The dual window slides past the speech material in a frame-by-frame fashion and the weighting of the inputs is automatically formed during the training of the phone network.

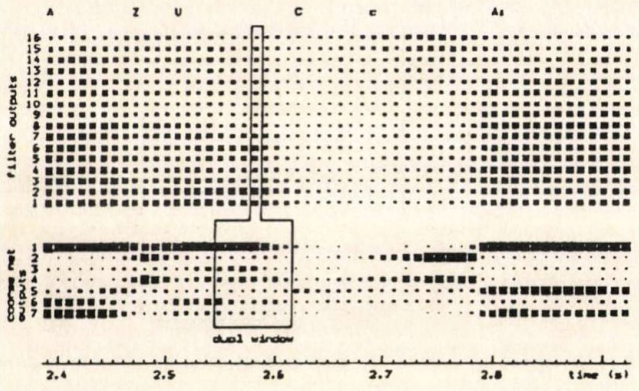


Fig. 4. The dual input window of the phone network. It includes one section of the filter bank output and the coarse phonetic feature parameters in seven frames.

Ideally, only the output node of the phone net associated with the actually processed phone should have a high activation, while all the others should remain at the base level. An example of a real output of the net is shown by representation 4 in Fig. 3. The sizes of the rectangles are proportional to the output activations of the phone network. It may be seen that only a few output nodes have high activation simultaneously. The most active node indicates the first phoneme candidate of the network, the next one corresponds to the second candidate, and so forth.

Automatic segmentation of speech is performed by a segmentation network. The activation levels of only the first candidates of the phone net are used as input, and the input is taken over a 15 frame wide window (150 ms). The single output of the network is expected to be zero except for an activation peak at the very first frame of each phoneme only. An example is shown in Fig. 3, representation 5. Phoneme labels for the detected segments are based upon an evaluation of the smoothed output activations of the phone net. This completes the transition from the time domain into the event domain, and the output can be seen at the bottom of Fig. 3.

Only the units D and H in Fig. 2 contain language specific elements. In units E and G, the size of the data memory needed depends on the size of phoneme set, but the function and the other parameters are independent of language. Changing language requires a reshaping of the network structure of unit D to the new phoneme set and retraining the unit with a new speech material. The phoneme table in unit H needs to be altered as well.

4. STRUCTURE, PRINCIPLES AND LEARNING OF DIFFERENT NETWORK ELEMENTS

4.1. Filter bank

The filter bank simulation program is based on an FFT-procedure [1] using a Hamming window. The FFT calculates 1024 spectrum lines, and their energy is merged within the filter bands.

4.2. The coarse feature classification neural network

This processing unit has multiple tasks. The most important one is to transfer spectral amplitudes into phonetically related features describing manner and place of articulation. A key problem is to select an appropriate set of coarse features. In spite of the existence of some different phonetic feature sets [6], [2], [3], [19], a new feature set was constructed according to the following principles:

- the feature parameters should be detectable within a single spectral frame
- the feature set should be language and speaker independent
- the feature set should describe the manner and place of articulation
- the feature set should be usable for different phonetic classes
- the feature target values should be readable from an existing speech database
- the total number of features should be small
- the features should conform to traditional phonetic feature sets in the steady state phase of the phonemes
- the feature set may be non-distinctive for phonemes, i.e., some phonemes may be identical in all features.

Table 2. Coarse phonetic features for the Hungarian phoneme set

	voiceness	noiseness	nasalness	frontness	centralness	backness	vowelness
. [pause]	-	-	-	-	-	-	-
B [b, occ.]	+	+	-	+	-	-	-
b [b, burst]	+	+	-	+	-	-	-
C [ts, occ]	-	-	-	+	-	-	-
c [ts, burst]	-	+	-	+	-	-	-
CS [ʧ, occ]	-	-	-	-	+	-	-
cs [ʧ, burst]	-	+	-	-	+	-	-
D [d, occ]	+	-	-	-	+	-	-
d [d, burst]	+	+	-	-	+	-	-
F [f]	-	+	-	+	-	-	-
G [g, occ]	+	-	-	-	-	+	-
g [g, burst]	+	+	-	-	-	+	-
GY [ɟ, occ]	+	-	-	-	-	+	-
gy [ɟ, burst]	+	+	-	-	-	+	-
H [h]	-	+	-	-	-	+	-
J [ʃ]	+	-	-	-	+	-	-
K [k, occ]	-	-	-	-	-	+	-
k [k, burst]	-	+	-	-	-	+	-
L [l]	+	-	-	-	+	-	-
M [m]	+	-	+	+	-	-	-
N [n]	+	-	+	-	+	-	-
NY [ɲ]	+	-	-	-	-	+	-
P [p, occ]	-	-	-	+	-	-	-
p [p, burst]	-	+	-	+	-	-	-
R [r]	+	-	-	-	+	-	-
S [ʃ]	-	+	-	-	+	-	-
SZ [s]	-	+	-	+	-	-	-
T [t, occ]	-	-	-	-	+	-	-
t [t, burst]	-	+	-	-	+	-	-
TY [c, occ]	-	-	-	-	+	-	-
ty [c, burst]	-	+	-	-	+	-	-
V [v]	+	+	-	+	-	-	-
Z [z]	+	+	-	+	-	-	-
ZS [ʒ]	+	+	-	-	+	-	-
A [ɔ]	+	-	-	-	-	+	+
A: [ɔ:]	+	-	-	-	+	-	+
E [ɛ]	+	-	-	-	+	-	+
E: [ɛ:]	+	-	-	+	-	-	+
I [i]	+	-	-	+	-	-	+
I: [i:]	+	-	-	+	-	-	+
O [o]	+	-	-	-	-	+	+
O: [o:]	+	-	-	-	-	+	+
W [ø]	+	-	-	-	+	-	+
W: [ø:]	+	-	-	-	+	-	+
U [u]	+	-	-	-	-	+	+
U: [u:]	+	-	-	-	-	+	+
Y [y]	+	-	-	+	-	-	+
Y: [y:]	+	-	-	+	-	-	+
*	+	-	-	-	+	-	+

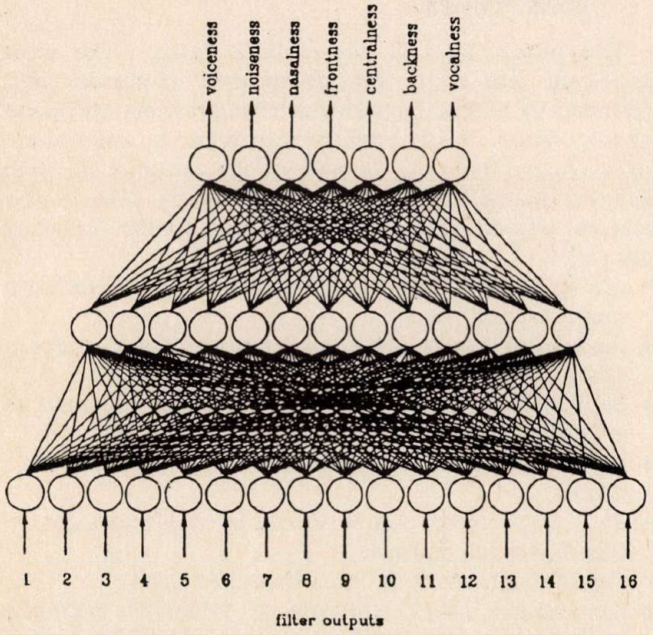


Fig. 5. The structure of the coarse feature network.

A fundamental problem is that the classical phonetic features are interpreted only at the phoneme level, and we need a description for 10 ms frames. This means that direct use of features is difficult, but there are still good reasons to use something related to them. At a phoneme border, two adjacent spectral frames may be very similar although they are labelled as different phonemes having different phonetic features. From an acoustical point of view, the difference may often be larger between the central frame and the boundary frames of a phoneme than between two adjacent frames on opposite sides of a phoneme boundary. The feature set is intended to capture some of these intra-phoneme variations, but the labels of speech data does not contain this information explicitly. However, the learning capacity of a neural network can be used for this purpose. The target values of the features are binary in the training phase, but by exposing the net to thousands of speech frames having varying acoustic representations for the same feature, we expect continuous feature values to develop that will convey also subtle details about each feature. This especially holds for the place of articulation related features.

As a result of a compromise among the principles listed above, a seven-element feature set was constructed based on manner and place of articulation related features. The manner related features are: voiceness, noiseness, nasalness and vowelness while the place related features are: frontness, centralness and backness, since these three can easily be assigned to both vowels and consonants. Since we expect the features to take continuous values, we use the suffix "-ness" is analogy with the term loudness for perceived sound level. Vowels have three features set positive: voiceness, vowelness and one place-related feature. Consonants are described by one place feature and the relevant manner features (vowelness is of course set to minus). Tables 1 and 2 show feature target values for Swedish and Hungarian phonemes, respectively. The features are not completely independent, and the targets are not able to discriminate between all the phonemes. However, this set is intended as a practical compromise.

A multi-layer perceptron net was trained using error back-propagation to recognize the coarse phonetic features from the spectral outputs of the filter bank. The network structure can be seen in Fig. 5. The net has 36 nodes (13 in the hidden layer) and 299 connections.

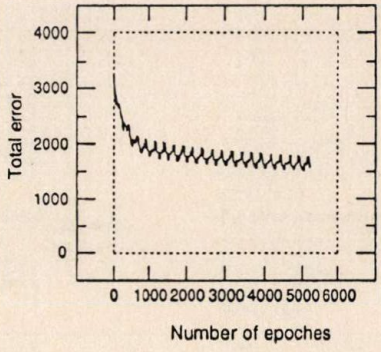


Fig. 6. The total error as a function of the number of epoches during the training of the coarse feature network. The fluctuations are caused by periodically changing the training material between five different training partitions. One epoch in this figure means one presentation of all the patterns in one partition only, and the error is also related to one partition at a time.

The fastest learning without getting disturbing oscillations in the error was achieved when using a learning rate of 0.001 and a momentum term of 0.9. The recognition results for the test set stabilized after a couple of thousand epoches. The total error as a function of the number of epoches is shown in Fig. 6. The training material was subdivided into five partitions, each fitting into the working memory of the computer. This speeded up the training by substantially reducing disk accesses. Each training fragment was used for a period of fifty epoches after which it was substituted by the next one. The fluctuation in the total error in Fig. 6 is due to this process.

4.3. The phone network

The phone network receives two types of inputs using a dual window (see Fig. 4). They are different both in content and in function. The first input is a seven frame window covering the coarse phonetic features. It contains some redundant information since during steady states, adjacent speech frames have similar feature values. Data compression can also be made by reducing the number of features. It is reasonable to compress these data before the final phoneme decision in order to reduce the network complexity and to decrease the number of weights to be trained. The compression is combined with a sort of time warping function to compensate for tempo changes in phoneme pronounciations. Both the compression and time compensation function are realized by a hidden layer with a special connection structure, as discussed, and the net should mix the two kinds of input data appropriately.

The number of output nodes in the network (the size of the poneme set) is also the result of a compromise. The result can be seen in Table 1 for the Swedish material and in Table 2 for the Hungarian material. A particular feature of both sets is to treat the occlusion and burst phase of stops and affricates as two different segments, where both segments have a specific label. Both phoneme sets have a special class for silent intervals marked by " ". These segments occur mostly at the end of the sentences. The vowels are represented by short and long pairs. The "v" symbol marks the neutral vowel.

According to the discussed requirements, the phone network consists of four layers: one input layer, one compression hidden layer, one "mixing" hidden layer, and one output layer. The network topology is shown in Fig. 7. As a result of separating the compression and mixing functions, the connection structure is quite complex. Each column of four nodes in the compression layer is connected to a group of three successive columns (frames) of the seven feature nodes, which are the output of the coarse feature net. There is a one-frame overlap in the coarse feature columns connected to the compression node columns. This means that all three compression columns will cover seven coarse feature columns. The rationale for this is that this will make them more insensitive to tempo changes in the speech.

The number of output nodes is determined by the phoneme inventory of the language being processed. The Swedish phone network has 40 output nodes and the Hungarian has 49. The number of nodes in the mixing layer is proportional to the size of the phoneme set. The mixing layer of the Swedish network has 20 nodes, the Hungarian 25. There are no differences in the lower level layers of the phone networks. The Swedish net has a total of 137 nodes and 1612 connections and the Hungarian has 151 nodes and 2177 connections. Fig. 7. shows only the first four and the last four nodes of the mixing and output layers.

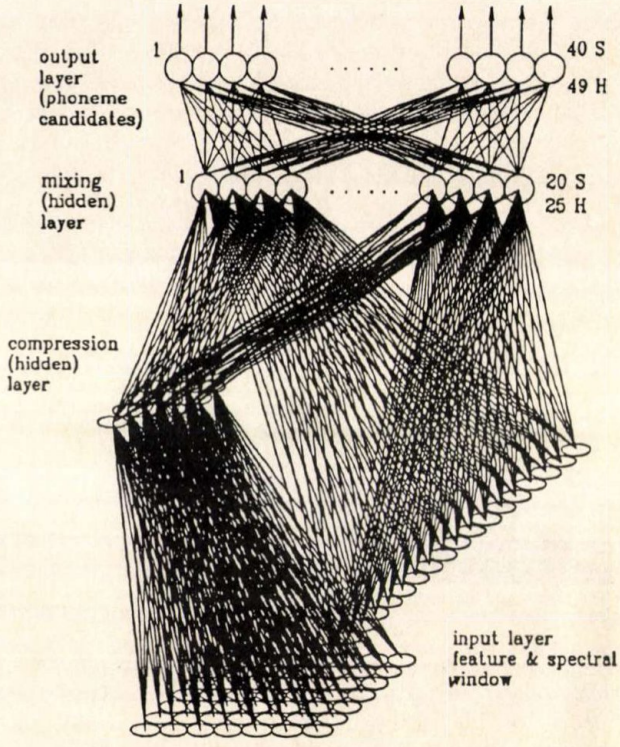


Fig. 7. The structure of the phone classification network. The Swedish network has 20 nodes in the mixing layer and 40 output nodes. The Hungarian network has 25 and 49 nodes, respectively.

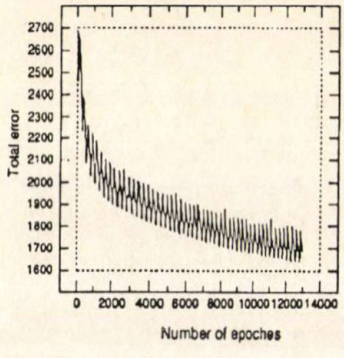


Fig. 8. The total error of the phone net as a function of the number of training epoches for the MAMO material. The fluctuations are caused by periodically changing the training material between five different training partitions. One epoch in this figure means one presentation of all the patterns in one partition only, and the error is also related to one partition at a time.

The fastest learning of the phone net was reached at a learning rate of 0.001 and a momentum term of 0.1. The weights were updated after each pattern. The training material was divided into the same five partitions used for the phone net in order speed up the training time

(the training time was still 100 hours CPU-time for the INTRED-material and 250 hours for the MAMO-material on an Apollo DN10000). As can be seen in Fig. 8, the total error has not reached a stable minimum value — the training process was interrupted due to running time limitations. On the other hand a low output error may be an indication of over-traninig so it is hard to know from this figure only whether the training time was too short. The oscillations are caused by the subdivision of the training set, as noted above. The number of "full" epoches for the training set is one fifth of the values in the Figure.

4.4. Segmentation neural network

The output activity of the phone classification network is shown as a surface in a three-dimensional space in Fig. 9.

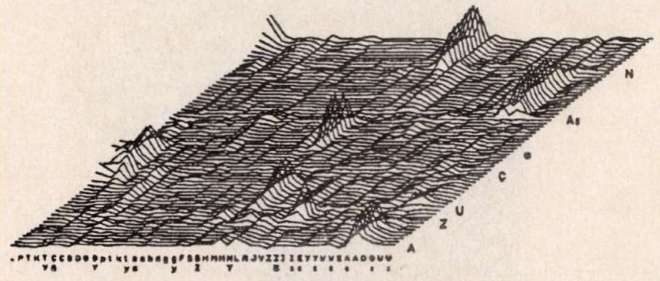


Fig. 9. The output activations of the phone network represented as a surface in a three-dimensional space: time, phoneme, set, and output activations. The manually decided phoneme labels are displayed along the time axis. Each ridge refers to one phoneme.

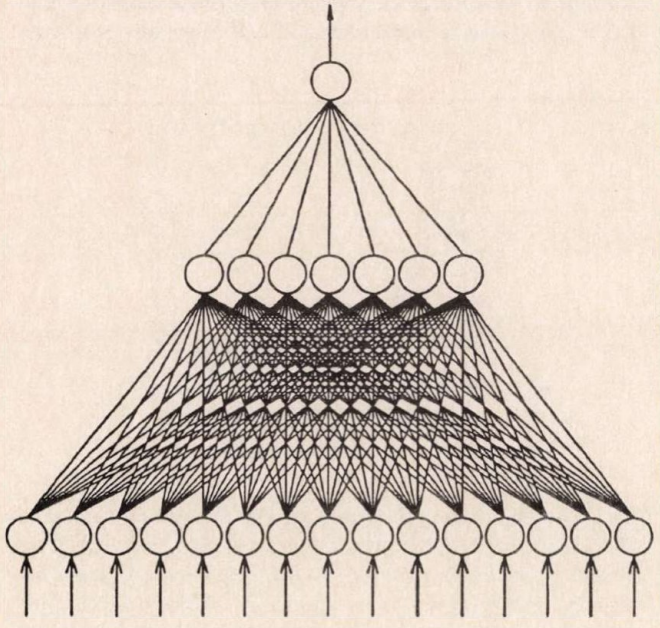


Fig. 10. The structure of the segmentation network with 15 input nodes, 7 hidden nodes and 1 output node.

The three axes are: the time, the phoneme set and the output activations. The target phoneme labels are marked along the time axis. The activation surface is typically flat with some "mountain ridges" running parallel to the time axes. Each ridge represents one phoneme. Multiple ridges in parallel indicate the existence of simultaneous phoneme candidates for that segment. The ridges are generally well separated in time and this implies that they might be used as a basis for segmentation. But it is not directly evident how to utilize them. The task to establish the decision criteria was again solved by a multi-layer network. The input parameter to the segmentation network is the amplitude of the phoneme output having the highest activation within each frame inside the input window that is 15 frames long (150 ms). The window is symmetric with seven frames on each side of the actual frame. The length of the window is selected to normally include at least one phoneme border. The segmentation network has a single output trained to show high activity (0.9) for the first frame of each phoneme and low activation (0.1) for all other frames — it is set up to "fire" at phoneme borders. The structure of the net is shown in Fig. 10. The training was done with a learning rate of 0.005, a momentum term of 0.1, and the weights were updated after each pattern.

4.5. Phone activation filter

It seems evident that the phoneme candidates for a segment should be selected among the candidates having the highest output activations. However, this will result in a lot of recognition errors due to spurious, short activation peaks. According to our experience, smoothing the activations by using a simple mean value filter gives significantly better recognition performance. The filter parameters were calculated emprically to produce optimal recognition. The first, second, and third phoneme candidates for each recognized segment were selected according to the level of their smoothed activation peaks. These phonemes constitute the output of the complete acoustic—phonetic level processing.

5. RESULTS

The speech material used for training was never used for testing the performance of the different networks. However, sometimes we give some results on the training set to indicate the generalization performance of the training.

5.1. Recognition of coarse phonetic features

A typical output of the coarse feature network may be seen in Fig. 11 where the output activations of the seven coarse features are plotted as a function of time. The segmentation and labelling was performed by a human phonetic expert. The straight horizontal lines indicate a 0.5 activation level. We notice steep transitions for the voiceness, noiseness and nasalness features as well as parts of the vowelness feature. These features almost behave like binary signals. The (nonlinear) network has become sensitive to the gradual filter bank variations and is usually able to follow the discrete target values quite reliably. Some exceptions to this can be found in Fig. 11 where the phoneme "Z" shows an uncertainty in the voiceness feature, may be due to strong frication, and in the phoneme "R" where the vowelness is somewhat ambigous, indicating its character of a semivowel. Most of the time feature transitions are in good synchronism with the manually set

boundaries. Features related to the place of articulation generally vary at a slower rate and these parameters frequently demonstrate intermediate values in spite of being binary during the training. The continuous distribution of these parameters agree with our original intentions since the articulation features covered by these parameters have a continuous character, and we expected the net to learn this by being exposed to numerous different varieties of these features during the network training. Some slow transitions for these parameters can be seen in Fig. 1, e.g., in the "AN"-sequence on the right hand side.

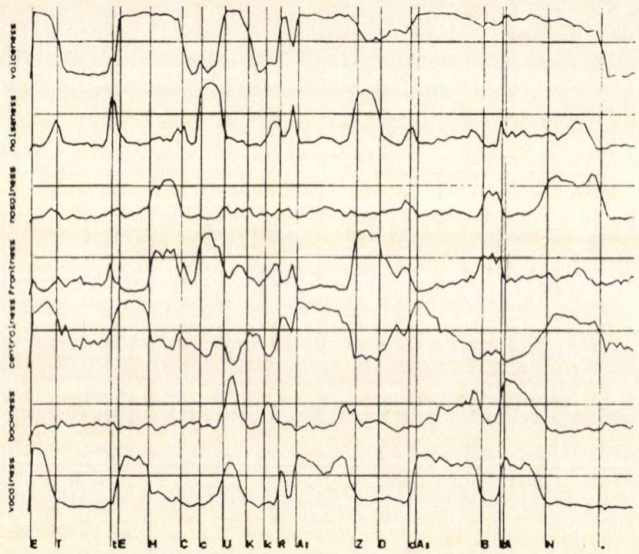


Fig. 11. Output activations of the coarse feature network as a function of time. Manually labelled segments. Horizontal lines indicate a 0.5 activation evel for each feature.

The performance of the coarse feature network was tested by using three different methods. In the most evident evaluation, a binary signal was formed by the output activations using a comparison level of 0.5 — if the signal exceeded this threshold the corresponding feature was set to one, and otherwise it was set to zero. These features were then compared with the features derived from the manual phoneme labels for each of the tested frames. This binary evaluation of the frame level recognition rate is summarized in Table 3. The network has a closely similar performance for both the INTRED- and the MAMO-material. Many features are correctly recognized for more than 90% of the frames, and all features are recognized above an 80% level. The manner of articulation-related features perform better than the place of articulation features. The last row gives the results when all features of a frame are correctly recognized. Most errors occur at phoneme transitions. The results show that in a substantial majority of the frames, the selected discrete phoneme features are detectable by the neural network at the frame level. Testing on the training set results in a 78.8% score for frames having all features correct which is just 2% higher than the results for the test set, indicating a good generalization for the coarse feature network.

Table 3. Performance of the coarse phonetic feature network on the frame level when evaluating the feature activations as binary signals

Feature	Percent correct feature recognition	
	INTRED	MAMO
voiceness	93.1	93.3
noiseness	91.0	92.9
nasalness	95.4	93.1
frontness	81.7	88.4
centralness	83.2	80.8
backness	88.7	88.2
vowelness	88.2	88.0
all features correct	76.9	80.0

5.2. Phoneme classification on the frame level

The phone network outputs an ordered list of phoneme candidates for each speech frame according to the level of their output activations — the best candidate has the highest activation. It should be noted that this decision is based on a relatively wide window (150 ms). Each frame has a target label selected by a phonetic expert making it easy to compare the assigned target phoneme to the network candidates. After the training process, 54.2% of the frames in the INTRED-test-material and 54.7% of the frames in the MAMO-material were recognized correctly. Testing on the training set for the INTRED-material resulted in 58.6% recognition, indicating a relatively good generalization also for the phone net, but also suggesting that the performance might be improved somewhat by continued training. It should be noted that all frames are included in these results — also transitional parts of the speech, where the acoustic character of the phonemes is changed due to coarticulation. The nature and the origin of the errors are treated in the next Section.

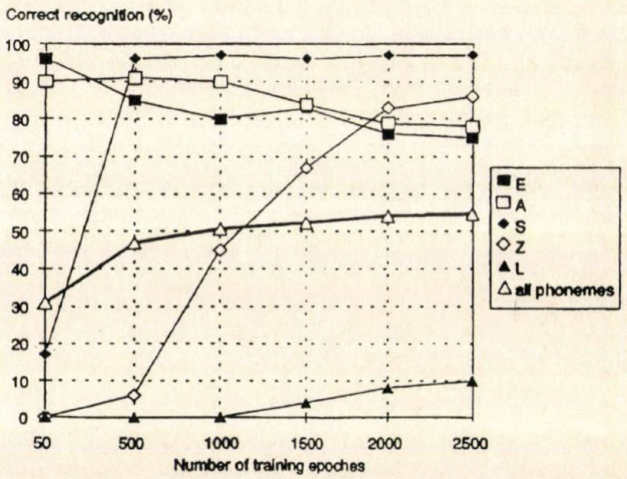


Fig. 12. The evolution of the phoneme recognition performance as a function of the training time for the Hungarian material. Each epoch means one presentation of all the training material.

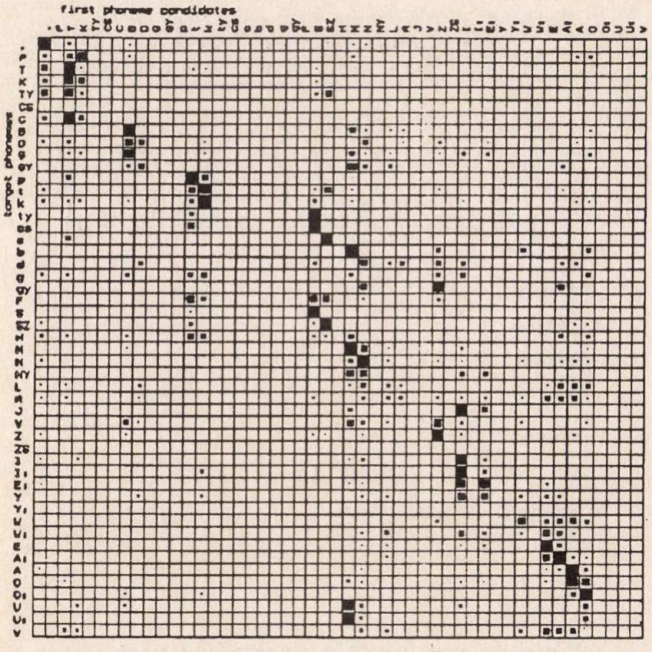


Fig. 13. Phoneme substitutions on the frame level based on the Hungarian MAMO-material. The values are normalized for each row. The area of the black squares is proportional to the nuimber of occurences in each position.

The training of the phone net was probably not completed when it was stopped due to time limitations. Some typical patterns can be seen in Fig. 12. The most frequent phoneme frames in the Hungarian material (E,A) have a very good recognition rate at the very beginning of the training process. However, not only the correct phonemes but almost all less frequent phonemes are classified as belonging to some of the most common ones in the beginning. As the net starts to differentiate among more phonemes, the performance becomes more and more sensitive to the low frequency phonemes. Their recognition rate starts to increase at different later phases of the training procedure and the increase has a different speed for different phonemes. The net result is that the overall recognition rate is monotonously increasing during the training time, though the rate is rather small at the end. It would, of course, be interesting to continue the training process, but it is very time consuming on our current facilities.

Phoneme substitutions are summarized by the confusion matrix in Fig. 13. The area of the black squares is proportional to the corresponding number of occurences. The sum of the values in each row is normalized and the sum over each row is a completely filled square. All of the most frequent confusions are caused by acoustic similarities.

5.3. Phoneme level classification using expert segmentation

This evaluation is based on the smoothed output activations of the phone network. An example of this is presented in Fig. 14. After observing all test material using this representation, the following significant conclusions could be drawn:

- The peak of the phone with the maximum activation has an average values of 0.47 inside the phoneme borders. The second and third largest peaks have a value or 0.27 and 0.20, respectively. Thus, the first candidate regularly has a considerably higher value than the others.
- The activation levels of adjacent phonemes cross each other close to the phoneme borders. This fact forms the basis for our automatic segmentation network.
- The phone candidates proposed by the net and the manual labels are the same in most cases. The recognition rate values are summarized in Table 4, in which each epoch includes one presentation of each pattern in the full training set.

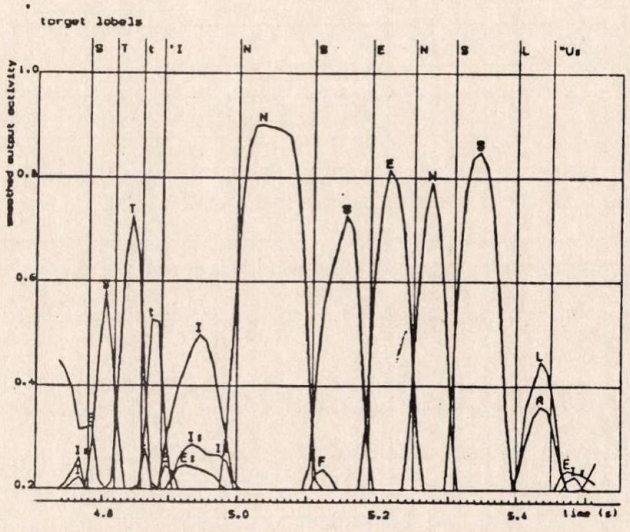


Fig. 14. An example of the smoothed phone net output activations.

The results for Hungarian show the evolution of the performance as a function of the phone net training. The performance shows some signs of levelling off at the end. The better scores for Swedish can probably be attributed to its smaller phoneme set.

Table 4. Phoneme recognition in percent for the Swedish and Hungarian phone network when using the segmentation of the human expert. Results for correct phoneme within the top one, two or three best phone candidates proposed by the phone net. Performance increase during training shown for the MAMO-network

Material	INTRED	MAMO					
No. of epoches	2500	50	500	1000	1500	2000	2500
1st corr.	64.4	23.2	42.2	46.7	46.5	59.6	50.4
1st or 2nd corr.	78.1	32.9	58.4	65.3	67.3	69.7	69.9
1st, 2nd or 3rd corr.	82.2	41.1	68.1	71.5	74.3	75.6	76.3

The peak value of the smoothed phoneme activations is a good measure of how reliable a network classification is. The relation between the recognition rate and the peak of the smoothed activation is shown in Fig. 15 showing that the probability of marking a correct classification is fairly well approximated by a linear function of the smoothed activation peak. However, sometimes wrong candidates were proposed even though the neural network had a high

activation value (0.7-0.8). This must, of course, be taken into consideration although one could argue that similar phenomena occur also in human communication.

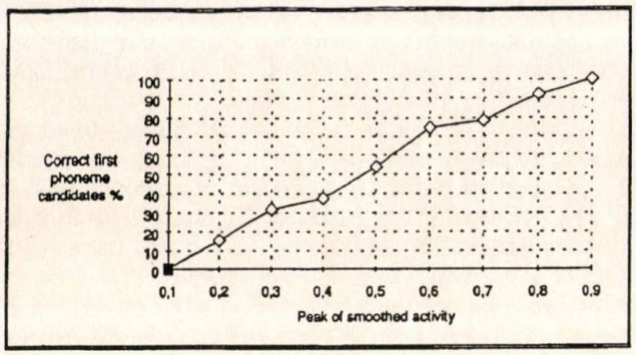


Fig. 15. The probability of correct phoneme recognition as a function of the smoothed activation peak values in the MAMO-material.

5.4. Recognition of phoneme borders

The single output of the segmentation network is trained to have a high value in the first frame of each phoneme, see signal form 5 in Fig. 3. The system detects a phoneme segment when the activation has a peak above the decision threshold. Increasing this threshold decreases the number of detected segments and vice versa. The optimum value can only be determined in cooperation with a language level processing. In our evaluation, a medium level threshold was chosen, where the number of errors did not vary very much when perturbing the threshold value. For the INTRED-material, a threshold of 0.2 was used and the value for the MAMO-material was 0.16. The results are summarized in Table 5.

Table 5. Performance of the segmentation network

	MAMO %	INTRED %
Segmentation in the frame marked by human	39.3	35.9
Segmentation within +/ − 1 frame	82.0	81.6
Lost segments	18.0	18.4
Extra segments	57.4	54.3

5.5. Phoneme recognition using automatic segmentation

The phoneme recognition process is based on the peaks of the smoothed phone net output activations within the segments detected by the segmentation net. The first three candidates were compared to the target phoneme in the evaluation. To reduce some of the problems due to lost and extra phoneme borders, phonemes in automatic segments placed plus or minus one (expert) segment from the correct target phoneme segment were allowed when comparing to the target phoneme. It is conjectured that these errors should not degrade an upper level recognition performance too much. Moreover, target segments were compared to phoneme candidates in multiple segments that overlapped them in time. Frequently, these inserted segments were correctly labelled which would reduce the degradation in recognition performance caused by them. Having these evaluation criteria in mind, we note that we get about the same level of performance as when using manual expert segmentation, see Table 6 below.

Table 6. Results for phoneme recognition when using automatic segmentation

	MAMO %	INTRED %
Correct first candidate	50.3	64.4
Correct first or second candidate	64.5	74.8
Correct first, second or third candidate	72.0	80.8

6. CONCLUSIONS

A phone-labelling system for continuous speech has been constructed and evaluated. Besides using an earlier established Swedish database, a Hungarian speech database has been assembled for training and testing the system.

The output activations of the network have an inherent ability to indicate parallel and overlapping speech events. They also have been shown to develop continuous values in spite of binary targets to discriminate between phones with similar target features, and in at least one case, these output values have a meaningful phonetic interpretation.

Many ideas regarding how to increase the system performance appeared during the evaluation of the system. The system is quite complex and has thousands of free parameters. Due to limited time, only a few parameter variations were tested. Below, some ideas are listed concerning parameters that either could improve the recognition rate or speed up the training process:

- the number of nodes in the hidden layers of the networks (coarse, phone, segmentation)
- the training parameters of networks (momentum term, learning rate)
- the selection of the training data (perhaps by representing rare phonemes more frequently than in natural speech and by omitting ambiguous elements)
- optimizing the smoothing filter parameters
- optimizing the feature window length
- optimizing the segmentation window length
- managing stops as single events in the phone classification.

The size of the speech material used is a limitation the effect of which we have some difficulties to estimate. Compared to the practically unlimited variations possible in a language, the representation of speech by 50 sentences, or 2800 phonemes, or 21000 frames, is very fragmentary,

but compared to similar systems reported in other papers, the size of the speech material is similar to many of them. There have been some speculations about the necessary acoustic-phonetic recognition accuracy in continuous speech recognition. It is a well known fact in telecommunication [4] that an 80% logatom intelligibility in a transmission system means practically error-free communication (99% sentence intelligibility). According to Klatt [11], some researchers argue that 60 to 70% accuracy is what we could expect from machines, while Klatt rather thinks that 90% is the needed performance target.

It is difficult to compare this system to other recently published systems. There are only a few results reported for complete phoneme sets, and the working principles, the speech materials used, and the evaluation methods are different. The speaker independent recognizer of AT&T has been reported to have a 52% phoneme recognition rate [12]. Systems based on the Kohonen feature map report a 75 to 90% recognition rate depending on the speaker [9]. Many system reporting recognition rates above 90% process a subset of phonemes only, or use presegmented phoneme samples. Considering that the above systems probably have been more elaborately tuned, we consider our efforts quite promising.

Future work includes evaluation of different input parameters, varying the sizes of input windows, speeding up the training and introducing a lexical component. It would also be interesting to test recurrent nodes and to do comparisons with self-organizing nets like those proposed by Kohonen, using the same speech material.

REFERENCES

[1] Bomberg, M., "Synthetic phioneme prototypes and source adaptation in a speech recognition system", *STL-QPSR*, No.1, pp.131–135.

[2] Chosmky, N. and Halle, M., *The Sound Pattern of English*, Harper & Row, Publ. New York.

[3] Fant, G., *Speech Sounds and Features*, The MIT Press, Cambridge, MA.

[4] Fletcher, H., *Speech and Hearing in Communication*, D. Van Nostrand Company, Princeton, NJ.

[5] Hunnicutt, S., "Acoustic correlates of redundancy and intelligibility", *STL-QPSR*, No.2-3, pp.7–14.

[6] Jacobson, R., Fant, G. and Halle, M., *Preliminaries to Speech Analysis. The Districtive Features and Their Correlates*, The MIT Press, Cambridge, MA.

[7] Kohonen, T., *Self-Organization and Associative Memory*, Springer-Verlag, Berlin.

[8] Kohonen, T., "The NEURAL' phonetic typewriter", *IEEE Computer* 3, pp.11–22.

[9] Kohonen, T., Torkkola, K. Shozakai, M. Kangas, J., and V'ant'a'a, O., "Microprocessor implementations of a large vocabulary speech recognizer and phonetic typewriter for Finnish and Japanese", pp.377–380 in *European Conf. on Speech Technology*, Edinburgh.

[10] Komori, Y., Hatazaki, K., Tanaka, T., Kawabata, T. and Shikano, K., "Phoneme recognition expert system using spectogram reading knowledge and neural networks", pp. 549–552 in (J. Tubach & J.J. Mariani, eds.) *Proc. Eurospeech 89, Paris, Vol.II.*, CPC Consultants Ltd, Edinburgh.

[11] Lea, W.E. (ed.), *Trends in Speech Recognition*, Prentice-Hall, Englewood Cliffs, NJ, p.266.

[12] Levinson, S.E., Liberman, M.Y., Ljolje, A. and Miller, L.G., "Speaker independent phonetic transcription of fluent speech for large vocabulary", pp.21–24 in *Proc. ICASSP – Glasgow.*

[13] Lippmann, R.P., "An introduction to computing with neural nets", *IEEE ASSP Magazine 4:2*, pp.4–22.

[14] Lippmann, R.P., "Neural nets for computing", pp.1–6 in *Proc. ICASSP – New York.*

[15] Mariani, J., "Recent advances in speech processing", pp. 429–440 in *Proc. ICASSP – Glasgow.*

[16] Niles, L., Silverman, H., Tajchman, G. and Bush, M., "How limited training data can allow a neural network to outperform an 'optimal' statistical classifier", pp.17–20 in *Proc. ICASSP – Glasgow.*

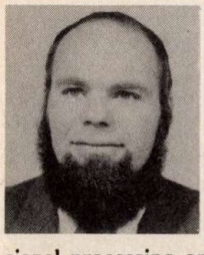
[17] Nord, L., "Acoustic–phonetic studies in a Swedish speech data bank", pp.1147–1152 in *Proc. SPEECH'88, Book 3* (7th FASE Symposium), Institute of Acoustics, Edinburgh.

[18] Rumelhart, D.E. and McClelland, J.E., *Parallel Distributed Processing, Vol.1-2*, The MIT Press Cambridge, MA.

[19] Singh, S., *Distinctive Features Theory and Validation*, University Park Press, Baltimore.

[20] Treleaven, P., "Neuroncomputers", *Int. J. Neurocomput.* 1:1, pp.4–31.

[21] Waibel, A., Hanazawa, T., Hinton, G., Shikano, K. and Lang, K., "Phoneme recognition using time-delay neural networks", *IEEE ASSP* 37:3, pp.626–631.



György Takács received the M.Sc. degree in telecommunication systems from the Technical University of Budapest, Hungary, in 1972, and Ph.D. in technical science from the Hungarian Academy of Sciences in 1991. Since 1972 he has been working at the PKI Telecommunications Institute. He is now the head of division for Strategic Development of Telecommunications. His interest is focused on speech signal processing and its application for new telecommunication services.

SPEECH QUALITY ASSESSMENT FOR LOW BIT-RATE CODING

S. MOLNÁR, P. TATAI and Z. JÁNOSY

DEPARTMENT OF TELECOMMUNICATIONS AND TELEMATICS
TECHNICAL UNIVERSITY OF BUDAPEST
H–1111 BUDAPEST, STOCZEK U. 2.

The paper reviews subjective and objective speech quality assessment methods. A particularly efficient method, the Paired Comparison Test based on a Modulated Noise Reference Unit (MNRU, CCITT Recommendation P.81) is described in detail. A speech quality test system called Qualiphon developed at DTT, TUB is also introduced. It is capable of both simulating and testing low bit-rate speech coders and various digital as well as analog communication channels. As an example of objective quality assessments using the built-in real time MNRU of the Qualiphon system, the procedure and some experimental results obtained with 4800 bit/s and a 3200 bit/s CELP coder are presented.

1. INTRODUCTION

Speech communication services represent a wide variety and steadily growing applications field for digital communication systems. The applications include public commercial telephone networks, mobile communications, satellite communications, private communication lines, switched networks, cellular telephones, voice storage services etc. From an economical point of view, a communication channel should transmit as much information as possible, therefore efficient speech digitization and compression methods are needed. These are particularly important in the case of radio communication systems and voice storage applications where the bandwidth and information capacity are severely limited, necessitating low bit-rate speech coding methods. In the following, low bit-rate means rates below the standard value of 64 Kbit/s.

All speech coding methods, however, have an undesirable side-effect, namely the degradation of speech quality. The fidelity of speech and the reduction of bit-rate are contradictory. The efficiency of speech digitization and compression can be measured directly by the resulting transmission bit-rate but the fidelity of speech cannot be interpreted easily because of its subjective nature. Some interpretations will be introduced in this paper. In order to evaluate speech coding systems, speech quality assessment methods are needed [1]. These are very important not only for optimizing coding algorithms but also for designing effective communication systems. There are two categories of such methods: subjective and objective speech quality assessments. The subjective methods are based on standardized procedures which use humans to judge the quality of speech. In contrast, the objective methods eliminate human judgements from the assessment procedure and provide computable results based on measurable physical quantities. The main problem of finding a good objective speech quality assessment method, however, is that its results should highly correlate with users' opinion, so once again one has to resort to subjective test in order to "calibrate" objective measures.

The main goal of this paper is to give a brief summary of subjective and objective speech quality assessment methods and to introduce the Qualiphon system which is an efficient speech quality assessing frame program based on the DSPLab signal processing environment developed at DTT, TUB.

2. SUBJECTIVE SPEECH QUALITY ASSESSMENT

Speech quality depends primarily on human perception so subjective quality assessment methods imply humans as referees. There are two categories of subjective measures: *utilitarian* and *analytic*. Utilitarian methods measure speech quality on a unidimensional scale so that results can be summarized by a single number capable of comparing communication systems directly. Analytic methods generate their results on a multidimensional scale reflecting various speech quality components.

2.1. Utilitarian methods

2.1.1. Intelligibility tests

There is a wide range of utilitarian methods used extensively mainly for very low bit-rate or synthetic voice. One category focused on speech intelligibility called *Intelligibility Tests* consists of articulation tests, rhyme tests and speech interference tests [2]. *Articulation Tests* [3] give their results as a percentage of correctly received sounds, words and sentences. The rate of correctly heard monosyllables is known as syllabic articulation. Sound articulation is defined as a percentage of correctly received phonemes. With modifying articulation tests the *Equivalent Loss Method* [3] is obtained. The score of this method is the Articulation Equivalent Loss (AEN) which is defined as the difference in attenuation values at 80% sound articulation between reference and test system.

2.1.2. Quality tests

Another category of utilitarian methods comprises *Quality Tests*. The intelligibility tests are unable to measure the speech quality when speech is highly intelligible, and this is exactly what we expect from most speech services. So methods are needed which can measure other attributes such as pleasantness or naturalness. For these purposes new methods have been developed. The most widely used method based on opinion rating is the *Mean Opinion Score* (MOS) [3]. In this test, five grades of speech quality are usually distinguished (Excellent, Good, Fair, Poor and

Bad) which are assigned by listeners. The measured quality is equal to the average value of scores received from all listeners.

In spite of being a good subjective measure of speech quality, MOS cannot distinguish fine differences between speech samples assessed by E-grade votes in case of high quality speech. To improve resolution at high quality level, the *Paired-Comparison Method* [4] has been developed. According to this method, two signals are presented to listeners who are asked to choose the better one. It is a forced comparison, "Equal" quality answer is not allowed. The percentage of the signal chosen as the preferred one is the preference score.

The *Equivalent Noise Paired Comparison Test* [4] is based upon a reference signal with varying signal-to-noise (SN) ratio. In this test, the quality is defined as the SN ratio of reference signal corresponding to 50% preference level. CCITT recommends a reference signal generator device called *Modulated Noise Reference Unit* (MNRU) for such tests in Recommendations P. 81 [5]. MNRU contains a white noise source modulated by the speech signal and the generated multiplicative noise is added to the speech signal to produce the reference signal (see Fig. 1). Such a reference signal has a speech component and a speech-amplitude correlated noise component with flat frequency spectrum. The signal-to-noise ratio (denoted by Q [dB]) can be set in the MNRU and is constant over the full dynamic range so its subjective effect is very similar to that of the distortion of logarithmic quantizers (standard PCM systems). Modifying MNRU with a noise-shaping filter results in a SN ratio which is almost independent from frequency.

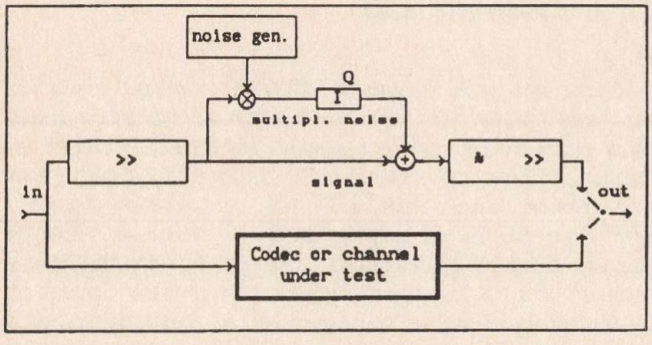


Fig. 1. Equivalent Noise Paired Comparison Test

Although it is sometimes difficult to compare signals with different types of impairments, the great advantage of the Paired Comparison Tests is that they provide highly accurate assessment on an absolute scale of speech quality even with as few as about 20 listeners. Therefore they are ideal for quick tests. The computer assisted speech quality assessment system Qualiphon developed at DTT, TUB and introduced in Sec. 4 is also based on the Equivalent Noise Paired Comparison Test. In contrary to this, quality tests based on MOS require hundreds or thousands of listeners to provide reliable results and the scale is relative, depending on time, country etc. like grades in schools. Nevertheless, thorough international investigations usually include MOS tests because it can take into account various kinds of deteriorations on a single scale.

2.2 Analytic methods

Analytic methods attempt to obtain different quality attributes of perceived speech by exploiting the phenomenon that listeners usually agree on the degree of speech impairment, but vary in their preference of that degradation. Therefore analytic methods generate a multidimensional characterization of the speech quality. Some methods have been developed which produce this kind of parametric description of speech such as *Paired Acceptability Rating Method (PARM), Quality Acceptance Rating Test (QUART) and Diagnostic Acceptability Measure (DAM)* [6].

For instance, in DAM a parametric scale is presented which is divided into three categories: signal category with ten rating scores, background category with seven rating scores and overall quality category with three rating scores. With the use of factor analysis, these twenty scores have been reduced to thirteen nearly independent perceptual quality scores. The signal category consists of the following parameters: fluttering, thinning, rasping, smothering, whining and irregularity. The background category consists of the following parameters: hissing, buzzing, bubbling and thumping. The general perceptual qualities are intelligibility, pleasantness and acceptability. From these parameters, overall attributes can be calculated such as total signal quality, total background quality and the most important factor which is based on all parameters, the composite acceptability.

Although the analytic methods provide a fairly good description of speech quality, such investigations are difficult and time consuming. No wonder that much effort has been paid to find objective speech quality measures which can efficiently predict subjective quality.

3. OBJECTIVE SPEECH QUALITY ASSESSMENT

However good assessment of speech quality can be provided by subjective tests, these have several disadvantages: they are expensive, slow, difficult to handle, non-repeatable due to the fact that human listeners' decisions depend on the test conditions and on their personal disposition. Especially the time-consuming nature of subjective measures excludes their use in the design and optimization of speech coding and communication systems.

Computable objective measures of speech quality based on measured physical parameters are much more desirable. They are cheap, simple, repeatable and fast in comparison with subjective measures, but they can be applied only if they predict subjective speech quality sufficiently well. So the task is to find an objective measure which can be efficiently computed from the original and distorted speech data set, and which highly correlates with subjective tests.

To solve this task is not easy because the human speech perception process is very complex and poorly understood. It involves also the grammar and other diverse factors such as the speakers' attitude and emotional state. People use a lot of redundant information in speech so as a result, certain slight distortion effects could cause complete intelligibility loss while other more extensive distortion products may be almost unperceivable. Quality assessment requires objective measures in order to take into consideration semantic, prosodic, syntactic, phonetic, etc. information. Of course, no objective measures

provide all these, and speech coding systems generally do not produce e.g. semantic distortions but only a fraction of all possible distortions. Accordingly, it is possible to find objective measures showing high correlation with subjective results.

3.1 Waveform distortion measures

Waveform distortion measures are defined in the time domain and are based on some kind of discrepancy between the original and the distorted speech waveform. These type of measures are known as variants of signal-to-noise ratio where noise is usually defined as the difference between the original and distorted signal. Owing to an inevitable coding or transmission delay, precise synchronization is necessary between the two waveforms.

3.1.1 Signal-to-Noise Ratio (SNR)

The conventional SNR is defined by Eq. (1), used for a long time:

$$\mathrm{SNR} = 10 \log_{10} \frac{\sum_{j=1}^{N} x(j)^2}{\sum_{j=1}^{N} [x(j) - y(j)]^2} \qquad (1)$$

where $x(j)$ and $y(j)$ denote the samples of the original input and the distorted output speech signals, respectively, and N is the number of speech samples considered. The correlation (R) of this measure with subjective measures ranges from 0.24 and somewhat higher values may be obtained using a multiple regression procedure [7].

3.1.2 Segmental Signal-to-Noise Ratio (SEGSNR)

As the conventional SNR is inadequate to predict subjective quality, the so-called segmental SNR has been proposed:

$$\mathrm{SEGSNR} = \frac{1}{\sqrt{M}} \sum_{i=1} \mathrm{SNR}_i \qquad (2)$$

where SNR_i is defined as in Eq. (1) in the signal frame i, and M is the number of frames. Here one frame is a segment of speech, usually 10 ... 30 ms long.

The SEGSNR is based on the experimental fact showing that the inherently nonstationary speech can be considered approximately stationary for the short interval of a frame. Since distortion effects depend on speech statistics, a measure which is the average of objective measures calculated separately for each frame provides a better quality indicator than overall measures.

The heuristic method using arithmetic mean of logarithmic quantities in (2) corresponds to equal weighting of high and low level sounds of an utterance. This is justified by the investigations resulting in high R values ranging from 0.77 to 0.95, where the upper limit is yielded only for one kind of waveform coder distortions [2].

Further developments in SNR have resulted in the *Frequency Variant Segmental SNR* ($R = 0.93$) [2] which takes account of the frequency distribution of distortion products. There are other variants, too, as the *Granular Segmental SNR, Articulation Index* and its improved method, the *Speech Transmission Index* [2].

These methods, however, can be applied only to waveform coders which attempt to reproduce the signal shape. More efficient coding algorithms exploit also the insensitivity of human perception to phase information and they reproduce waveforms that show little resemblance to the original speech. In this case, the distortion products can no longer be separated by a simple subtraction in the time domain.

3.2 Spectral distortion measures

These measures are defined in frequency domain between the original and distorted speech spectra. The *Spectral Distortion (SD)* defined by Eq. (3) provides a logarithmic spectral distortion measure which can be computed by FFT:

$$\mathrm{SD} = \left[\frac{1}{\pi} \int_0^{\pi} \left[\ln \frac{S_x(\omega)}{S_y(\omega)} \right]^2 d\omega \right]^{\frac{1}{2}} \qquad (3)$$

$S_x(\omega)$ and $S_y(\omega)$ denote the original and distorted speech spectrum, respectively. Experiments have yielded $R = 0.6$ for SD [2].

3.3 Spectral envelope distortion measures

These measures are based on the spectrum envelope distortion computed by Linear Predictive Coding (LPC). They attempt to provide macroscopic (smoothed) spectral information which is believed to be the most important speech signal characteristic.

3.3.1 Log Likelihood Ratio (LR)

This measure expresses the dissimilarity between all-pole models of the original and distorted speech waveforms. It is defined by the Equation:

$$\mathrm{LR} = \ln \left[\frac{1}{\pi} \int_0^{\pi} \left| \frac{A_y(e^{j\omega})}{A_x(e^{j\omega})} \right|^2 d\omega \right] = \ln \left[\frac{\overline{a}_y R_x \overline{a}_y^T}{\overline{a}_x R_x \overline{a}_x^T} \right] \qquad (4)$$

where $A_x(z)$ and $A_y(z)$ are the analysis filters given by $\overline{a}_x$ and $\overline{a}_y$ LPC coefficient vectors of the original and distorted speech, respectively, and R_x is the autocorrelation matrix of the original speech. The elements of R_x are defined as follows:

$$r_x(|i-j|) = \sum_{n=1}^{N-|i-j|} x(n)x(n+|i-j|) \qquad (5)$$
$$\text{for} \quad |i-j| = 0, 1, 2, \ldots, p.$$

where N is the length of the frame and p is the prediction order used in LPC. A simple interpretation of LR can be given with the aid of analysis and synthesis filters of the LPC [8].

The correlation of LR with subjective measures is $R = 0.59$ [2]. Other popular variants of LR are the *Itakura-Saito Measure*, the *COSH Measure* and the *Weighted Log Itakura-Saito Measure* [2], which provide similar performance but originate from a somewhat different mathematical reasoning.

3.2.2 LPC Cepstrum Distance Measure (CD)

The CD measure is based on cepstral coefficients ($c(k)$) [9,10] computed by LPC:

$$\mathrm{CD} = \left[[c_x(0) - c_y(0)]^2 + 2 \sum_{k=1}^{L} [c_x(k) - c_y(k)]^2 \right]^{\frac{1}{2}} \quad (6)$$

Using Parseval's relation, it can be seen that CD is equivalent to a log spectral distance of cepstrally smoothed spectra. For CD the correlation with subjective results has ranged from 0.8 to 0.9 [3,4].

3.3.3 Other LPC Measures

There exist also other measures such as *Feedback Coefficient Distance Measures* ($R = 0.06 - 0.11$), *PARCOR Coefficient Distance Measures* ($R = 0.11 - 0.46$) and *Area Ratio Distance Measures* ($R = 0.24 - 0.62$) [2]. These are similar but not so successful measures as the LR and CD measures described above.

3.4 Auditory distortion measures

The task of objective measures is to predict the human perception effects. This goal cannot be reached without modeling the human perception mechanism. Some measures discussed earlier attempt to give such models (frequency weighted measures divide the frequency domain into bands according to the critical bands of human hearing, log differences are used according to the Fechner's law, etc.), but the hearing models behind these measures are oversimplified. Recently, researches in this field have attempted to construct better and more sophisticated models. The *Bark Spectral Distortion (BSD)* measure is based on the Bark spectrum, which reflects the ear's nonlinear scale of frequency and amplitude. The Bark spectrum is derived from the classic speech spectrum computed by FFT applying three additional transformations. It executes a critical-band transformation which is a frequency scale warping according to the human hearing, a pre-emphasis which simulates the equal loudness curves of the ear and an intensity-loudness conversion. The Bark Distance Measure is defined in Eq. (7) [11].

$$\mathrm{BSD}(k) = \sum_{i=1}^{N} [S_x(k,i) - S_y(k,i)]^2 \quad (7)$$

where $S_x(k,i)$ and $S_y(k,i)$ are the Bark spectrums of original and distorted speech for the k^{th} segment and i^{th} band, and N is the number of bands. The overall distortion is then the average BSD over all frames. $R = 0.85 - 0.98$ is obtained for BSD [11].

3.5 Composite measures

Another approach is to combine different objective measures using multiple regression analysis to maximize the correlation with subjective results. Although some improvements have been obtained in this way and the correlation coefficients have varied from 0.8 to 0.996 (the upper limit was found only for a restrictive class of waveform coders), the multidimensional scaling of objective measures has shown that the various objective measures are not independent. This means that even though these measures are arithmetically quite dissimilar they all measure practically the same features of speech, hence no significant improvement can be achieved by combining them into one composite measure.

A more promising approach is to design objective measures which predict individual qualities of speech and to combine them into a new measure called *Parametric Objective Measure* [2]. Parameters like those used for DAM seem suitable because they indicate perceptually different aspects of speech quality. Recently such a measure has been developed with good correlation coefficient ($R = 0.82$) over broad classes of distortions [2].

The short overview above illustrates that the design of an appropriate objective measure is not a trivial task. Since the speech generation and perception mechanisms are very complex, no simple objective measure can ever be expected. In fact, objective measures can be designed only for a limited and well defined class of distortions. The fast progress in speech coding, processing and transmission systems, however, necessitates a continuous research to cope with the ever increasing variety of degradations.

This kind of research could well be supported by two important tools:

a) a signal processing simulation frame program in order to investigate speech coding algorithms and various objective distortion measures,
b) a subjective speech quality assessment system for convenient and fast testing.

In the following, the Qualiphon system is described which has been developed to fulfill these tasks.

4. THE QUALIPHON SYSTEM

The Qualiphon system is an integrated subjective speech quality assessment system which is based on the DSPLab integrated development environment [12].

4.1 The DSPLab development environment

The DSPLab system is a toolbox and a graphical environment for simplifying the development of digital signal processing (DSP) algorithms. The toolbox handles the recording and playback of speech samples, the file I/O, the graphic display and user interface elements. The integrated environment makes it easy to record, store, display and manipulate the signals by providing a graphical user interface, using menus and dialogues, and a two-channel oscilloscope for a graphic display of the signals.

The main feature of this system is the support of user defined processing algorithms. The environment is provided as a library which can be linked to the user's program. The user program can contain several procedures that perform some transformations on a signal. These user defined procedures will appear in the menu

structure of the integrated environment, so in fact they become a part of the environment.

The two-channel oscilloscope can be used to compare the input and output of the algorithm visually (as shown in Fig. 2) as well as by listening to the samples, a feature which is especially useful for speech coded simulations. The environment supports the subjective and objective measurements by providing reference signal generators such as sinusoidal and pulse waves, and a Modulated Noise Reference Unit (MNRU) recommended by the CCITT for subjective listening tests. At present, only two types of signal to noise ratio (SNR) measurements are included, but more sophisticated objective measures are in development. The DSPLab system is written entirely in Borland's Turbo Pascal and it supports several DSP boards, like the MegaMicro TMS32010 board or the Ariel DSP-16, so it is possible to use DSP machine code algorithms too. The necessary communication routines are provided by the toolbox.

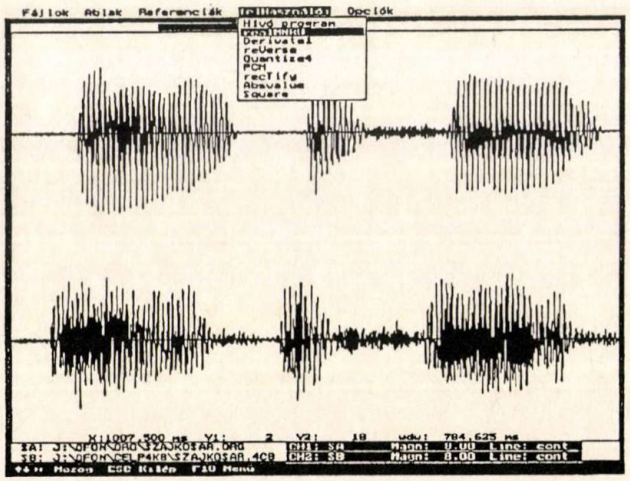


Fig. 2. An example of DSPLab oscilloscope function

4.2 System description

Investigations on subjective speech quality assessment methods have yielded the conclusion that Equivalent Noise Paired Comparison Test recommended by CCITT [5] provides a relatively fast and reliable way of subjective testing. In the process of such tests, the preparations, the presentation of speech material and the evaluation of results are the most time consuming tasks. Because the listening and decision making time cannot be reduced, the purpose was to make the preparations and evaluations efficient. The computer assisted *Qualiphon* system [1] performs the following tasks: collecting speech samples, generating reference sentences, presenting the speech materials to the listening process, evaluating results etc.

In this way, Qualiphon is an integrated subjective speech quality assessment system which automates the process of subjective listening test using an MNRU (according to the CCITT Recommendation P. 81.). The system has been developed for evaluating the subjective quality of low bit-rate speech codecs but it can be extended to include components needed for testing communication channels.

Qualiphon has a speech sample library which can be used for recording the reference sentences. The samples to be evaluated can be recorded from the output of a hardware codec, or can be the results of software simulations (e.g. using the DSLab system).

A test editor is provided to put together the sentence pairs and to specify the MNRU values used during the test. Five pairs are used for each test, and the order of the MNRU values and the sentences of the pairs to be used as reference can be randomized for minimizing sentence dependence.

The system uses a signal processor for implementing a real-time MNRU. During the test, five pairs of sentences are presented, and after each pair the listener has to decide which one was "better". The results are accumulated for each of the test series, and after a sufficient number of results have been collected from the listeners, an evaluation can be performed automatically. The distribution of decisions is shown on a graph, and the equivalent subjective SNR (Q) is calculated.

4.3 Examples of subjective speech quality assessment

Several speech coding algorithms have already been tested by the Qualiphon system. One of the promising algorithms at low bit-rates, 8 Kbit/s and below, is the CELP (Code Exited Linear Prediction) coding. As an example, the testing of a real time CELP codec at rate of 4800 and 3200 bit/s is presented in the following.

4.3.1 The CELP codec [13]

The CELP codec selected for subjective testing is realized by a DSP in real time. It conforms with the Proposed US Federal Standard 1016, jointly developed by the US DoD and AT&T Bell Laboratories for 4800 bit/s voice coders.

In the CELP coder, a Linear Prediction (LP) analysis is applied to estimate a 10^{th} order LP filter by means of the autocorrelation method, using a 30ms Hamming window. The results of the LP analysis, the linear prediction coefficients are transformed into Line Spectrum Pairs (LSP) coefficients. The optimum excitation sequences for the LP synthesis filter are obtained from an adaptive and a stochastic codebook, thus the CELP algorithm can be considered as a two-stage vector quantizer for the speech signal. The algorithm includes a joint optimization process for code vector index and the quantized gain factor.

4.3.2 The test procedure

For subjective speech quality assessment, the Equivalent Noise Paired Comparison Test is applied with MNRU, using the Qualiphon system.

The speech material for the test consisted of two sets of sentences. Each set contained 5 pairs of sentences. In each sentence pair there was a distorted (CELP) sentence, and a reference sentence with a predetermined level of multiplied noise. The order of distorted and reference

sentence and the changing of distortion level were randomized. In order to eliminate sentence dependency, in the second set of sentence-pairs the roles of reference and test sentences were exchanged. The number of listeners was 25. The test procedure was carried out with both 4800 and 3200 bit/s CELP codecs.

4.3.3 Experimental results

The result of preference evaluation for a 4800 bit/s CELP codec is shown in Fig. 3. In this figure, the 50% preference point is indicated which gives the Q value of the tested codec. After averaging all tests, the Q value is 11.7 dB for 3200 bit/s, and 17.1 dB for 4800 bit/s.

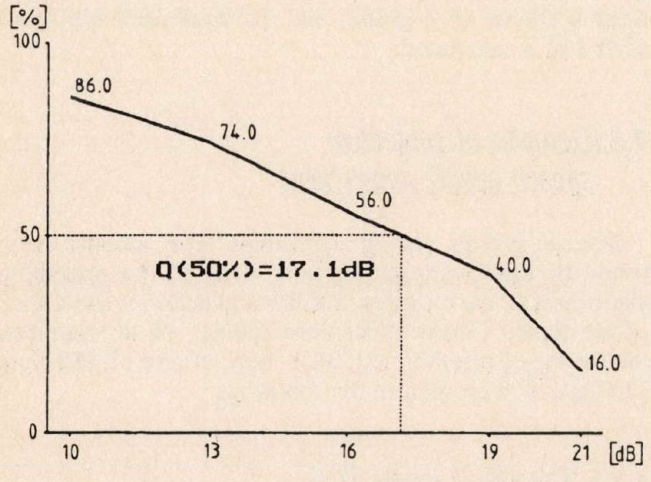


Fig. 3. The result for preference test for a 4800 bit/s CELP coder

The acceptable limit for commercial telephone service is about $Q = 20$dB. According to CCITT Recommendation G. 113. [14], this is about equivalent to the performance of 14 asynchronously tandemed 8-bit PCM codecs, i.e. 14 qdu (quantizing distortion units) based on a $15 \log_{10} n$ law, where n is the number of qdu's [5].

As can be seen, neither 3200 nor 4800 bit/s CELP codecs meet the requirements. Based on preliminary tests, however, it is expected that the 7200 bit/s CELP codec (not yet available in hardware form) can provide acceptable speech quality.

5. CONCLUSION

The paper presents a wide variety of subjective and objective speech quality assessment methods, and discusses the main difficulties arising due to the subjective nature of the problem. It has been found that one of the most efficient methods is the Paired Comparison Test based on a Modulated Noise Reference Unit (MNRU) which has already been standardized by the CCITT.

In order to simplify the time consuming and laborious preparation and evaluation work of subjective testing, the speech quality test system Qualiphon has been developed at DTT, TUB. It contains a built-in real time MNRU for convenient testing. The Qualiphon system is based on a DSPLab integrated digital signal processing environment which is also suitable for simulating low bit-rate speech coders and various digital as well as analog communication channels.

Finally, experimental results of the subjective testing of a 4800 bit/s and a 3200 bit/s CELP coder are presented. The results evaluated by the Qualiphon system indicate that quality of these codecs is not satisfactory for commercial telephone applications. According to preliminary tests, the same codec at a bit-rate 7200 bit/s provides acceptable quality.

In the present form, the Qualiphon system can compute only simple objective measures as e.g. SNR. Further research work is needed to develop and realize more reliable objective speech quality measures.

ACKNOWLEDGEMENTS

Thanks are due to Dr. G. Gordos, Head of DTT-TUB, for his continuous interest and support during the research work.

REFERENCES

[1] G. Gordos, P. Tatai et. al., "On digitized speech quality assessment", OMFB Tanulmányok, 1987–1990 (in Hungarian)

[2] S. R. Quackenbush, T. P. Barnwell, III and M.A. Clements, "Objective Measures of Speech Quality", Prentice Hall, 1988.

[3] N. Kitawaki, M. Honda and K. Itoh, "Speech Quality Assessment Methods for Speech-Coding Systems", *IEEE Communications Magazine*, Vol.. 22, no. 10, pp. 26–33, Oct. 1984.

[4] N. Kitawaki and H. Nagabuchi, "Quality Assessment of Speech Coding and Speech Synthesis Systems", *IEEE Communications Magazine*, Vol.. 26, no. 10, pp. 36–44, Oct. 1988.

[5] CCITT "Recommendations of the P Series", Red Book, Vol.. V. VIIIth Plenary Assembly, 1984, and Blue Book, Vol.. V. IXth Plenary Assembly, 1988.

[6] W. D. Voiers, "Diagnostic Acceptability Measure for Speech Communication Systems", *IEEE Proc. Int. Conf. Acoust., Speech Signal Processing*, pp. 204–7, May. 1977.

[7] B. J. McDermott, C. Scagliola and D. J. Goodman, "Perceptual and Objective Evaluation of Speech Processed by Adaptive Differential PCM", *Bell Syst. Tech. J., Vol..57,* No. 5, pp.1597–1618, June 1978

[8] P. Tatai, "Comments on Objective Quality Measures in Speech Encoding", *Budavox Review, No. 4,* pp. 20–24, 1989

[9] Gordos G., Takács Gy., Digital Speech Processing, Műszaki Kiadó, Budapest, 1983 (in Hungarian)

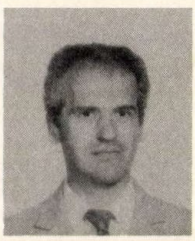
[10] L. Hanzó, L. Hinsenkamp, " On the Subjective and Objective Evaluation of Speech Coders", *Budavox Review, No. 2,* pp. 6–8, 1987

[11] S. Wang, A, Sekey and A. Gresho, "Auditory Distortion Measure for Speech Coding", *IEEE Proc. Int. Conf. Acoust., Speech, Signal Processing* pp. 493–496, 1991.

[12] Z. Jánosy: Integrated Environment for the Development of Digital Signal Processing Algorithms, TDK Pályázat, BME-HEI, Nov. 1989 (in Hungarian)
[13] L. A. Hernández-Gómez et al., "Real-Time Implementation and Evaluation of Variable Rate CELP Coders", *IEEE Proc. Int. Conf. Acoust., Speech, Signal Processing*, pp. 585–588, 1991.
[14] CCITT, "Recommendations G.101-G.181", Blue Book, Vol.. III. 1, IXth Plenary Assembly, 1988

Sándor Molnár graduated in 1991 at the Technical University of Budapest, Faculty of Electrical Engineering, in Telecommunications. He won the Republic's Scholarship two times and the first prize of the Frigyes Csáki Automation Competition in 1990. His diploma work was the software and hardware design of a supervisory system for a PCM network. Since September 1991 he is working for his Ph. D. degree at the Department of Telecommunications and Telematics. His principal interest is speech processing, including speech coding, recognition and synthesis. At present he is researching in the field of speech quality measures.

Zoltán Jánosy was admitted to the Faculty of Electrical Engineering, Technical University of Budapest in 1986. Since 1988, he has been working part time at the Department of Telecommunications and Telematics as a student. He has been writing programs for digital speech coding, speech quality assessment and graphical environment for signal processing and computer music. His thesis for the diploma is about computer processing of player piano rolls.



Péter Tatai graduated in 1964 at the Faculty of Electrical Engineering, Technical University of Budapest, in Telecommunications. From 1964 to 1986 he was employed by the Research Institute for Telecommunications, Budapest, where he was involved in the research and development of communication equipment and automatic test systems. In 1976 he became the Head of Code Modulation Systems Department and was active in the development of the Hungarian PCM system. He has studied telecommunications and digital signal processing abroad at the University of Tokyo, at Imperial College, London and at the Royal Institute of Technology, Stockholm. Since 1986 he is with the Department of Telecommunications and Telematics, Technical University of Budapest. He gives lectures both in Hungarian and English on telecommunications. Most of his time, however, is devoted to research, including the guidance of several undergraduate and graduate students. His present research interest includes digital signal processing in general, speech quality assessment, telecommunication testing, voice coding, synchronous digital networks etc. He has more than 30 publications.

EFFICIENT SEARCH IN DISSIMILARITY SPACES FOR AUTOMATIC SPEECH RECOGNITION

A. FARAGÓ, T. LINDER and G. LUGOSI

DEPARTMENT OF TELECOMMUNICATIONS AND TELEMATICS
TECHNICAL UNIVERSITY OF BUDAPEST
H-1111 Budapest, Stoczek u. 2.

It is a fundamental problem in automatic speech recognition to find quickly the nearest neighbor of a given point among a large number of other points in a high dimensional space, where the distance is defined by a complicated dissimilarity measure. We investigate here a family of algorithm that work very efficiently in practice but no exact analysis has been published to establish their properties. We prove in a general probabilistic model that asimptotically these algorithms find the nearest neighbor by a constant number of distance computations.

1. INTRODUCTION

It is known that in Euclidean spaces the nearest neighbor of a query point among n other sample points can be found in $O(1)$ time, at the price of preprocessing ([1], [2], [3]). This is a problem one encounters quite often in a number of applications, such as pattern recognition, vector quantization of signals, computational geometry etc. In some situations, however, the underlying space is more complicated: no explicit coordinate structure is given and the "distance" between the points is merely a measure of dissimilarity but not a metric in the usual sense [9]. This is the case e.g. in automatic speech recognition and in other fields of pattern recognition where measuring the similarity of objects is a complicated issue. In such a harder situation the so called cell or bucketing techniques of the above cited references do not work, as they require coordinate structure. Also, these methods become impractical very rapidly even in Euclidean spaces with increasing dimension.

Several attempts have been made to work out fast algorithms for these more difficult instances ([1], [2], [3]). Their efficiency has been shown by extensive computer simulation but not with mathematical analysis. In the present paper we prove in a general probabilistic model that an algorithm of this type finds the nearest neighbor by $O(1)$ distance computations on the average.

2. THE ALGORITHM

In order to establish a sufficiently general setting we introduce the concept of a dissimilarity space, which is a generalization of the notion of metric space.

Definition 1. A nonempty set D with a function $\rho : D \times D \to R$ is called a dissimilarity space if for any $x, y \in D$ the following conditions are satisfied:

$$\rho(x,y) \geq 0,$$
$$\rho(x,y) = 0 \quad \text{iff} \quad x = y,$$
$$\rho(x,y) = \rho(y,x).$$

A dissimilarity space in which the triangle inequality holds is a metric space. Just as in metric spaces, a subset H of a dissimilarity space is called bounded if $\sup\{\rho(x,y) : x, y \in H\} < \infty$. The notation of the metric is relaxed here but (obviously) one needs to impose some geometric structure and dimensionality on a dissimilarity space.

Definition 2. Let D be a dissimilarity space and let $\alpha \geq \beta > 0$. The points $z_1, z_2, \ldots, z_k \in D$ are said to form a basis at level (α, β) for a set $H \subset D$, if for any $x, y \in H$

$$\alpha\rho(x,y) \geq \max_{i=1,\ldots,k} |\rho(x,z_i) - \rho(y,z_i)| \geq \beta\rho(x,y) \tag{1}$$

is satisfied. Moreover, a dissimilarity space D is called finite dimensional if there exist $\alpha \geq \beta > 0$ and a positive integer k (depending on D only) such that for any bounded subset $H \subset D$ there are k points in D which form a basis at level (α, β) for H.

Example 1. It is not hard to see that $\mathcal{R}^d$ with the Euclidean metric is a finite (e.g. $d + 1$) dimensional dissimilarity space. A possible basis for a bounded set $H \subset \mathcal{R}^d$ is formed, for example, by the vertices of a sufficiently large regular d-dimensional simplex containing H. Elementary geometric calculations show that level values $\alpha = 1$ and $\beta = 1/2$ can be chosen.

Example 2. Let P be a full dimensional bounded polytope in $\mathcal{R}^d$ with vertices $z_1, z_2, \ldots, z_k$. Denote by $\alpha_i(x,y)$ the angle under which the line segment $\overline{xy}$ is seen from z_i. Set

$$\rho(x,y) = \max_i \alpha_i(x,y).$$

It is left to the reader that the points of P with dissimilarity measure ρ is a finite dimensional dissimilarity space. Another infinite family of examples is given by the following result whose proof is omitted here.

Theorem 1. Every finite dimensional normed vector space is a finite dimensional dissimilarity space with dissimilarity measure $\rho(x,y) = \|x - y\|$.

The nearest neighbor searching problem in a dissimilarity space is the following. We are given a set of n points $X_1, \ldots, X_n$, elements of a bounded set $H \subset D$. A nearest neighbor algorithm should determine in an efficient way, using some preprocessing of the points, the closest of these points to a new query point X coming from H. Here closeness means similarity, that is, the nearest neighbor of X is X_i if $\rho(X, X_i) \leq \rho(X, X_j), j = 1, \ldots, n$. The common idea of the (coordinate free) algorithms [4], [5], [6], [7], [8] is that they restrict the search to some appropriately chosen neighborhood of the query point with the following crucial properties:

- The neighborhood is large enough to guarantee to contain the nearest neighbor.
- The neighborhood is small enough to ensure that the average number of sample points contained remains asymptotically bounded.
- The neighborhood is defined constructively via distances to such points which are already known in the preprocessing stage.

To grasp and analyze this common idea we declare an algorithm which contains it in a pure form isolated from additional factors.

Let D be a finite dimensional dissimilarity space and let the points $z_1, \ldots, z_k$ form a basis for the bounded set H at level (α, β). Our proposed algorithm is the following.

Algorithm 1.

Preprocessing. Compute and store all the values $\rho(X_i, z_j)$, $i = 1, \ldots, n$; $j = 1, \ldots, k$. As k is fixed, this means $O(n)$ preprocessing time and storage cost.

Nearest neighbor searching
INITIALIZATION: Set $\mathcal{T} \leftarrow \{X_1, \ldots, X_n\}$.

STEP 1: Compute the value of

$$\gamma(X_i) = \max_{j=1,\ldots,k} |\rho(X_i, z_j) - \rho(X, z_j)|$$

for each $X_i \in \mathcal{T}$.

STEP 2: Set $t_0 \leftarrow \min_i \gamma(X_i)$. Delete all the points X_l from $\mathcal{T}$ for which

$$\gamma(X_l) > \frac{\alpha}{\beta} t_0$$

holds.

STEP 3: Find the nearest neighbor of X in the remaining part of $\mathcal{T}$ by exhaustive search:

$$T^{NN} = \operatorname*{argmin}_{U \in \mathcal{T}} \rho(X, U)$$

STOP, T^{NN} is the result.

The next theorem shows the correctness of the algorithm.

Theorem 2. Algorithm 1 always finds the nearest neighbor.

Proof. We have to show that the correct nearest neighbor, which we denote by X_n^{NN}, is never deleted from $\mathcal{T}$ in Step 2. Set

$$X_n^* = \operatorname*{argmin}_{i=1,\ldots,n} \gamma(X_i).$$

In the definition of X_n^{NN} and X_n^*, in case of ambiguity, we choose a random index among the candidates. Assume indirectly that $\gamma(X_n^{NN}) > \frac{\alpha}{\beta}\gamma(X_n^*)$, that is, Step 2 excludes X_n^{NN}. But from this, using Definition 2, we have

$$\rho(X, X_n^{NN}) \geq \frac{1}{\alpha}\gamma(X_n^{NN}) > \frac{1}{\beta}\gamma(X_n^*) \geq \rho(X, X_n^*),$$

a contradiction.

It is quite clear that in the worst case, that is, when no exclusion is carried out in Step 2, the algorithm executes n dissimilarity calculations. However, the next section shows that in a rather general probabilistic setup the average case is substantially different from the worst case. In particular, the number of dissimilarity calculations remains constant on the average as n increases.

3. PROBABILISTIC ANALYSIS

For the analysis of the average complexity we have to set up a probabilistic model. Let $(D, \mathcal{S})$ be a measurable space, where the family of sets $\mathcal{S}$ is termed the collection of measurable subsets of D. It is assumed that the measurable sets of the finite dimensional dissimilarity space D include the closed balls $B(x, r) = \{y \in D : \rho(x, y) \leq r\}$ of radius r centered at x for all $r > 0, x \in D$. We assume furthermore that $\rho : D \times D \rightarrow \mathcal{R}$ is a Borel measurable function on the product measurable space $(D \times D, \mathcal{S} \times \mathcal{S})$. Note that in the examples mentioned above these conditions are satisfied.

Let $X, X_1, \ldots, X_n$ be independent identically distributed random elements taking their values from a bounded subset H of D. Introduce the notation

$$p(x, r) = P_X(B(x, r)) = \Pr\{X \in B(x, r)\}.$$

We assume that the following regularity condition holds for the common distribution of $X, X_1, \ldots, X_n$.

Condition 1. There exist a $d > 0$ and a function $f : D \rightarrow \mathcal{R}$ such that

$$\lim_{r \to 0} \frac{p(x, r)}{r^d} = f(x) > 0 \qquad (2)$$

uniformly for almost all $x \in D (mod P_X)$.

Now we can state the main result.

Theorem 3. Let F_n be the number of dissimilarity calculations executed by Algorithm 1 for n points. If Condition 1 holds, then

$$\limsup_{n \to \infty} E(F_n) \leq k + \left(\frac{\alpha}{\beta}\right)^{2d},$$

where $E(\cdot)$ denotes expectation and k, α, β are as in the description of the algorithm.

Before proving the theorem rigorously it is worth mentioning that the main idea is the following: we show that a ball of radius $c\rho(X, X_n^{NN}), c > 0$ fixed, centered at the query point X contains asymptotically only a constant number of sample points, on the average. To present the exact proof we need a nontrivial property of finite dimensional dissimilarity spaces.

Lemma 1. Let X be a random element taking its values from a finite dimensional dissimilarity space D. Suppose that $\Pr\{X \in A\} = 1$ for some bounded measurable subset A of D. Then for any fixed $r_1 > 0$ there exists an $\epsilon > 0$ such that

$$\Pr\{p(X, r_1) \geq \epsilon\} = 1.$$

As the proof of Lemma 1 is quite involved we have to omit it.

Proof of Theorem 3. As the $\rho(X_i, z_j)$ values are given by the preprocessing, therefore, Step 1 of the algorithm requires only k dissimilarity calculations. Thus it is enough to consider the number of points T_n not deleted from $\mathcal{T}$ in Step 2 for $F_n = k + T_n$. Let X_n^* and X_n^{NN} be as in

the proof of Theorem 2. Using Definition 2, for each X_i which remains in $\mathcal{T}$ after Step 2, we have

$$\begin{aligned}\rho(X,X_i) &\le \frac{1}{\beta}\gamma(X_i) \le \frac{\alpha}{\beta^2}\gamma(X_n^*) \\ &\le \frac{\alpha}{\beta^2}\gamma(X_n^{NN}) \le \frac{\alpha^2}{\beta^2}\rho(X,X_n^{NN}).\end{aligned} \tag{3}$$

Put $c = \frac{\alpha^2}{\beta^2} \ge 1$. Denoting by T_n' the number of X_i with $\rho(X,X_i) \le c\rho(X,X_n^{NN})$, by (3) we have $T_n \le T_n'$, thus it suffices to show that

$$\lim_{n\to\infty} E(T_n') = c^d, \tag{4}$$

which is exactly what we will do. From now on in the proof I_B will denote the indicator of the set B and the short notation $R_n = \rho(X,X_n^{NN})$ will be used. Now, using the i.i.d. property of $X, X_1, \ldots, X_n$, we can write

$$\begin{aligned}E(T_n') &= E\left(\sum_{i=1}^{n} I_{\{X_i \in B(X,cR_n)\}}\right) \\ &= nE(I_{\{X_n \in B(X,cR_n)\}}) \\ &= nE(I_{\{X_n \in B(X,cR_n)\}} I\{X_n = X_n^{NN}\}) \\ &\quad + nE(I_{\{X_n \in B(X,cR_n)\}} I_{\{X_n \ne X_n^{NN}\}}).\end{aligned} \tag{5}$$

The first term in (5) is obviously $n\frac{1}{n}$ while the second can be written as

$$\begin{aligned}&nE(I_{\{X_n \in B(X,cR_{n-1})\}}) - \\ &\quad nE(I_{\{X_n \in B(X,R_{n-1})\}}),\end{aligned} \tag{6}$$

where the second term is again $n\frac{1}{n}$. Thus (5) amounts to

$$\begin{aligned}E(T_n') &= nE(I_{\{X_n \in B(X,cR_{n-1})\}}) \\ &= nE(E[I_{\{X_n \in B(X,cR_{n-1})\}}|X]), \\ &= nE[p(X,cR_{n-1})]\end{aligned}$$

where in the last step we used the independence of the X_i. Since $E[p(X,R_{n-1})] = \Pr\{X_n \in B(X,R_{n-1})\} = \Pr\{X_n = X_n^{NN}\} = \frac{1}{n}$, we conclude that for $r > 0$

$$\begin{aligned}E(T_{n+1}') &= \frac{E[p(X,cR_n)]}{E[p(X,R_n)]} \\ &= \frac{E[p(X,cR_n)I_{\{R_n \le r\}}] + E[p(X,cR_n)I_{\{R_n > r\}}]}{E[p(X,R_n)I_{\{R_n \le r\}}] + E[p(X,cR_n)I_{\{R_n > r\}}]}.\end{aligned}$$

Now, by Lemma 1, $\epsilon > 0$ can be chosen that

$$\Pr\{R_n > 1\} = E[(1-p(X,r))^n] \le (1-\epsilon)^n,$$

that is, $\Pr\{R_n > r\}$ tends to zero exponentially fast. Since the second terms in both the numerator and the denominator of (6) are upper bounded by this probability and since the denominator is $\frac{1}{n}$ and the numerator is greater, it follows that

$$\lim_{n\to\infty} E(T_{n+1}') = \lim_{n\to\infty} \frac{E[p(X,cR_n)I_{\{R_n \le r\}}]}{E[p(X,R_n)I_{\{R_n \le r\}}]} \tag{7}$$

for arbitrary $r > 0$ provided that the limit on the right-hand side exists. But, by the uniform convergence in Condition 1, for any $\epsilon > 0$ an $r > 0$ can be chosen such that the following inequalities hold

$$\begin{aligned}&\frac{E[1-\epsilon)f(X)(cR_n)^d I_{\{R_n \le r\}}]}{E[(1+\epsilon)f(X)R_n^d I_{\{R_n \le r\}}]} \\ &\le \frac{E[p(X,cR_n)I_{\{R_n \le r\}}]}{E[p(X,R_n)I_{\{R_n \le r\}}]} \\ &\le \frac{E[(1+\epsilon)f(X)(cR_n)^d I_{\{R_n \le r\}}]}{E[(1-\epsilon)f(X)R_n^d I_{\{R_n \le r\}}]}.\end{aligned}$$

After cancellations we obtain

$$\frac{1-\epsilon}{1+\epsilon}c^d \le \frac{E[p(X,cR_n)I_{\{R_n \le r\}}]}{E[p(X,R_n)I_{\{R_n \le r\}}]} \le \frac{1+\epsilon}{1-\epsilon}c^d. \tag{8}$$

Since ϵ is arbitrary (7) and (8) together imply

$$\lim_{n\to\infty} E(T_n') = c^d,$$

and the proof is completed.

4. CONCLUSION

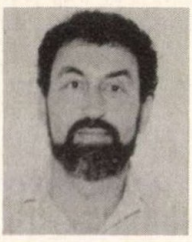
The algorithm and its analysis should be considered as an attempt to find the mathematical foundations of a family of fast nearest neighbor algorithms working well in practice in high dimensions, under general conditions, using no coordinates of the sample points. As a measure of complexity the number of dissimilarity ("distance") calculations has been chosen ignoring all the side computations. This point of view can be defended considering the following facts. Firstly: the practical simulation results in the cited refereces show that the running time of the algorithm is essentially determined by the number of dissimilarity computations. Secondly: the side computations in Step 2 of the Algorithm actually mean that one has to execute a full search in a transformed space where any $Y \in D$ is represented by the k-tuple $\tilde{Y} = (\rho(Y,z_1),\ldots,\rho(Y,z_k))$ and the distance is induced by the maximum norm. However, this problem is simpler than the original one and it is possible to use the existing cell technique solutions of low complexity (for a survey see [3]). Therefore, what the number of dissimilarity calculations means is the additional complexity induced by the more general instance. So the results can be interpreted such that finding the nearest neighbor in these more general spaces is theoretically of the same complexity as doing so in Euclidean spaces. On the other hand the new algorithmic idea is necessary, as cell/bucketing methods cannot be implemented efficiently for the general problem.

REFERENCES

[1] D.Dobkin and R.J.Lipton, "Multidimensional searching problems," *SIAM J.Comput.*, Vol.5.2, pp. 181-186, June 1976.

[2] J.H.Friedman, J.L. Bentley, and R.A. Finkel, "An algorithm for finding best matches in logarithmic expected time," *ACM*

Trans. Mathematical Software, Vol. 3.3, pp. 209-226, Sept. 1977.
[3] J.L. Bentley, B. W. Weide, and A. C. Yao, "Optimal expected-time algorithms for closest point problems," *ACM Trans. Mathematical Software,* Vol. 6.4, pp. 563-580, Dec. 1980.
[4] I. K. Sethi, "A fast algorithm for recognizing nearest neighbors," *IEEE Trans. Syst., Man, Cyber.,* Vol. SMC-11 pp. 245-248, March 1981.
[5] E. Vidal, "An algorithm for finding nearest neighbors in (approximately) constant average time," *Patt. Recogn. Lett.,* Vol. 4.3, pp. 145-157, July 1986.
[6] A. Faragó, T. Linder, G. Lugosi, and T. Pikler, "On the algorithmic problems of the nearest neighbor method" *Híradástechnika (Telecommunications),* pp. 337-341, Aug. 1988. (in Hungarian).
[7] K.Motoishi and T. Misumi, "Fast vector quantization algorithm by using an adaptive search technique," presented at IEEE Int. Symp. Inform. Theory, San Diego, CA Jan. 14-19, 1990.
[8] T. Linder and G.Lugosi, "Classification with a low complexity nearest neighbor algorithm," presented at IEEE Int. Symp. Inform. Theory, San Diego, CA Jan. 14-19, 1990.
[9] G. Gordos, "New Feature Extraction Methods and the Concept of Time-Warped Distance in Speech Processing," GLOBECOM'91, Phoenix, Arizona, December 1991, pp. 725-729.



András Faragó graduated at the Technical University of Budapest in 1976 and obtained his Ph.D. in electrical engineering in 1981, after doing graduate studies at the Virginia Polytechnic Institute in the USA. After graduation he joined to the Department of Mathematics at the Faculty of Electrical Engineering. In 1982 he moved to the Institute of Communication Electronics and now he is working as an associate professor at the Department of Telecommunications and Telematics of the Technical University of Budapest. His professional interest is in several fields: automatic speech recognition, telecommunication networks and he is also doing research in discrete optimization. He has just returned from a sabbatical year spent in the USA at the University of Massachusetts at Amherst, doing research in the area of telecommunication networks.

Gábor Lugosi received the M.S. and Ph.D. degrees in electrical engineering from the Technical University of Budapest and the Hungarian Academy of Sciences in 1987 and 1991, respectively. He is currently an Assistent Professor at the Department of Mathematics, Faculty of Electrical Engineering of the Technical University of Budapest, Hungary. His main interests are statistical pattern recognition, speech recognition, information theory and nonparametric statistics.

Tamás Linder received the M.S. degree in 1988, and the PH.D. degree in electrical engineering from the Hungarian Academy of Sciences in 1992. Since the fall of 1991 he has been with the Informatics and Electronics Research Group of the Hungarian Academy of Sciences. He is a member of the IEEE. His research interest includes information theory, source coding and quantization, as well as statistical pattern recognition.

PARAMETER ESTIMATION OF HIDDEN MARKOV PROCESSES WITH APPLICATION IN ISOLATED WORD RECOGNITION

A. FARAGÓ and G. LUGOSI

DEPARTMENT OF TELECOMMUNICATIONS AND TELEMATICS
TECHNICAL UNIVERSITY OF BUDAPEST
H-1111 Budapest, Stoczek u. 2.

Hidden Markov processes are very useful for describing the statistical properties of time-varying signals. For Hidden Markov models the estimation of the describing parameters is a basic problem. The paper presents a simple algorithm which provides a fast solution of the estimation problem in a practically important case.

1. INTRODUCTION

Hidden Markov processes have been found to be very useful for description of statistical properties of time varying signals. This type of characterization can be applied efficiently in economy, ecology and cryptanalysis, but its most frequent application is modeling human speech for automatic speech recognition tasks. The central problem concerning Hidden Markov models is the estimation of the describing parameters (which determine all the finite dimensional distributions of the process) from a given finite observation sequence.

In this paper we describe a simple algorithm which provides a fast solution of the estimation problem in a practically important case. The obtained parameters are optimal in a certain sense.

2. HIDDEN MARKOV PROCESSES

First we give a definition of a Hidden Markov process: Let $S = (S_1, S_2, \ldots)$ be a finite-state Markov chain with state set $Q = (q_1, q_2, \ldots, q_N)$, state-transition probability matrix $A = \{a_{ij}\}_{N\times N}$ and initial distribution vector $\Pi = (\pi_1, \pi_2, \ldots, \pi_N)$. (It means that $P(S_{t+1} = q_j | S_t = q_i) = a_{ij}$ and $P(S_1 = q_i) = \pi_i$ for $i, j = 1, 2, \ldots, N$, $t = 1, 2, \ldots$.) However, the process is "invisible", that is, all we can observe is the stochastic sequence $X = (X_1, X_2, \ldots)$, which satisfies $P(X_t = v_k | S_t = q_j) = b_{jk}$, where $B = \{b_{jk}\}_{N\times M}$ is a stochastic matrix, $V = (v_1, v_2, \ldots, v_M)$ denotes the output alphabet. Informally, a probability distribution $(b_{j1}, b_{j2}, \ldots, b_{jM})$ is associated with each state $q_j (j = 1, 2, \ldots, N)$, and whenever the underlying (hidden) Markov chain S is assumed to be in state q_j, the process emits an output symbol according to the corresponding probability distribution. (We can treat absolute continuous output distributions too, but for the sake of simplicity we consider finite output set.) The statistical description of such a process is given by a triple $M = (\Pi, A, B)$, which we call *model*.

In isolated word recognition each word (i) is represented by a model $M^{(i)}$, and every utterance of a given word is considered to be a realization of the stochastic process described by the distributions of $M^{(i)}$. If $X = (X_1, X_2, \ldots, X_T)$ is the spoken word to be recognized, then we choose the model for which $P_{M^{(i)}}(X)$ is maximal, that is, for which the probability of the occurrence of the given utterance is maximal. Alternatively, $P_{M^{(i)}}(X, S^*)$ can be used to be maximized, where S^* is the optimal state sequence, for which $P_{M^{(i)}}(X, S)$ is maximal.

3. ESTIMATION OF THE MODEL PARAMETERS

The crucial problem of modeling with hidden Markov processes is how to determine the parameters of the model M, when a finite sequence of the observation process $X = (X_1, X_2, \ldots, X_T)$ is given. It is proved [1], that in case of ergodic processes the maximum likelihood estimation is asymptotically consistent, that is, it gives the correct values with probability one, if the length of the observed sequence tends to infinity. Hence, it seems to be a good strategy to find the maximum likelihood estimation. Unfortunately, there is no known algorithm for this purpose. The existing methods (see e.g. [2], [3]) provide only convergence with uncertain speed to a local optimum of the likelihood function $P_M(X)$.

However, the situation can be changed by a slight modification of the objective function: Choose the so called "state optimized joint likelihood" $P_M(X, S^*)$ as the function to be maximized (where

$$S^* = \operatorname*{argmax}_S P_M(X, S),$$

as before). In this case we can give a simple dynamic programming algorithm, which results in the global optimum of the new objective function for the special, but important case of hidden Markov processes, namely, for "left-to-right models", where the underlying Markov chain has a following property: if the process leaves a state, it will never return to it (with probability one).

To describe the algorithm, we introduce some notations. For $i < j$ $X^{(ij)}$ denotes the subsequence $X_i X_{i+1} \ldots X_j$ of X. We write $v_k \in X^{(ij)}$ if the symbol v_k occurs in $X^{(ij)}$. Denote by $v_k // X^{(ij)}$ the relative frequency of occurrences of v_k in $X^{(ij)}$, that is, the number of occurrences of v_k in $X^{(ij)}$ divided by the length of $X^{(ij)}$, which is $j - i + 1$. Finally, let $V // X^{(ij)}$ be defined

by

$$V//X^{(ij)} = \prod_{v_k \in X^{(ij)}} v_k//X^{(ij)}$$

where the product is taken only over those v_k's, which occur in $X^{(ij)}$, so the factors of zero value are omitted.

Now we are prepared to describe the algorithm:

Step 1.

Construct a trellis with $N-1$ columns and with T nodes in each column. Denote by v_{ij} the jth node in the ith column ($i = 1,\ldots,N-1$; $j = 1,\ldots,T$). Add two additional nodes v_{00} and v_{NT} to the trellis, so $i = j = 0$ and $i = N$, $j = T$ will be allowed in the algorithm, as well. Draw an arc from v_{ij} to $v_{i+1,k}$ iff $i \le j \le T-N+i$, $j+1 \le k \le T-N+i$ and $0 \le i \le N-1$ hold. To the arcs assign the following weights:

$$w(v_{ij}, v_{i+1,k}) = c(j,k) + \log(V//X^{(j+1,k)})$$

where

$$c(j,k) =$$

$$= \begin{cases} \log\left[(1-\frac{1}{k-j+1})^{k-j+1} \cdot \frac{1}{k-j+1}\right] & \text{if } i \le N-2 \\ 0 & \text{if } i = N-1 \end{cases}$$

Step 2.

Run the Viterbi algorithm on the trellis constructed in Step 1. to find a maximum weight path from v_{00} to v_{NT}. The Viterbi algorithm is given here by the recurrence

$$\phi(v_{00}) = 0$$

$$\phi(v_{i,k}) = \max_j[\phi(v_{i-1,j}) + w(v_{i-1,j}, v_{ik})]$$

$$(i = 1,\ldots,N)$$

The maximum is taken over all j's, for which the arc $(v_{i-1,j}, v_{i,k})$ exists. We also have to keep record the nodes of the maximizing path, denote it by

$$v_{0,l_0}, v_{1,l_1}, v_{2,l_2}, \ldots, v_{i,l_i}, \ldots, v_{N,l_N}$$

$$(l_0 = 0, l_N = T)$$

Step 3.

Define the model parameters by

$$p_i = \frac{1}{l_i - l_{i-1}+1} \qquad (i = 1,\ldots,N-1)$$

$$p_N = 0$$

$$b_i(k) = v_k//X^{(l_{i-1}+1,l_i)} \qquad (i = 1,\ldots,N)$$

(where the l_i's have been obtained in Step 2.)

The proof of the optimality can be found in [4].

Unfortunately, we can not prove the asymptotic consistency of this type of estimation, but in speech processing tasks it is well motivated and gives good results.

REFERENCES

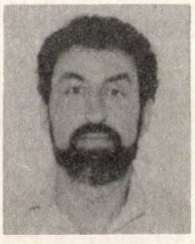
[1] Baum, L.E., Eagon, J.A., "An Inequality with Applications to Statistical Estimation for Probabilistic Functions of Markov Processes and to a Model for Ecology," Bull.AMS, 73 (1967) pp.36-363.

[2] Baum, L.E., Petrie, T., Soules, G., Weiss, N., "A Maximization Technique Occuring in the Statistical Analysis of Probabilistic Functions of Markov Chains," Ann. Math. Stat., 41 (1970) pp. 164-171.

[3] Levinson, S.E., Rabiner, L.R., Sondhi, M. M., "An Introduction to the Application of the Theory of Probabilistic Functions of a Markov Process to Automatic Speech Recognition", Bell System Technical Journal, No.4, pp. 1035-1074, April 1983.

[4] Faragó A., Lugosi G.: "A Fast Algorithm to Find the Global Optimum of Left-to-Right Hidden Markov Model Parameters", Problems of Control and Information Theory, 18 (1989), pp. 435-444.

[5] Gordos G., "New Feature Extraction Methods and the Concept of Time-Warped Distance in Speech Processing", GLOBECOM'91, Phoenix, Arizona, December 1991, pp. 725-729.



András Faragó graduated at the Technical University of Budapest in 1976 and obtained his Ph.D. in electrical engineering in 1981, after doing graduate studies at the Virginia Polytechnic Institute in the USA. After graduation he joined to the Department of Mathematics at the Faculty of Electrical Engineering. In 1982 he moved to the Institute of Communication Electronics and now he is working as an associate professor at the Department of Telecommunications and Telematics of the Technical University of Budapest. His professional interest is in several fields: automatic speech recognition, telecommunication networks and he is also doing research in discrete optimization. He has just returned from a sabbatical year spent in the USA at the University of Massachusetts at Amherst, doing research in the area of telecommunication networks.

Gábor Lugosi received the M.S. and Ph.D. degrees in electrical engineering from the Technical University of Budapest and the Hungarian Academy of Sciences in 1987 and 1991, respectively.

He is currently an Assistant Professor at the Department of Mathematics, Faculty of Electrical Engineering at the Technical University of Budapest, Hungary. His main interests are statistical pattern recognition, speech recognition, information theory and nonparametric statistics.

THE INTRINSIC BIMODALITY OF SPEECH COMMUNICATION AND THE SYNTHESIS OF TALKING FACES*

CHRISTIAN BENOIT

INSTITUT DE LA COMMUNICATION PARLÉE,
UNITÉ DE RECHERCHE ASSOCIÉTE AU CNRS N 368 INPN/ENSERG
UNIVERSITÉ STENDHAL, BP 25X
F38040 GRENOBLE, FRANCE

1. INTRODUCTION

In 1989, Negroponte predicted that "the emphasis in user interfaces will shift from the direct manipulation of objects on a virtual desktop to the delegation of tasks to three-dimensional, intelligent agents parading across our desks", and that "these agents will be rendered holographically, and we will communicate with them using many channels, *including speech and non-speech audio, gesture and facial expressions*" [62].

Historically, the talking machine with a human face has been a mystical means to power for charlants and shamans. In that vein, the first *speaking robots* were probably the famous statues in ancient Greek temples, whose power as oracles derived from a simple acoustic tube! The statues were inanimate, even though their impresionable listeners attributed a soul (*anima*) to them, because of their supposed speech competence. If this simple illustion already made them seem alive, how much more powerful would it have been if statue's faces were animated?

One can only wonder how children would perceive Walt Disney's or Tex Avery's cartoon characters if their facial movements were truly coherent with what they are meant to say, or with its dubbing into another language. Of course, these imaginary characters are given so many extraordinary behavioral qualities that we easily forgive their peculiar mouth gestures. We have even become accustomed to ignoring the asynchrony between Mickey's mouth and his voice.

What about natural speech? When a Candidate for the Presidency of the United States of America exclaims "Read my lips!", he is not asking his constituency to lip-read, he is simply using a classical English formula so that his audience must belive him, as it was written on his lips: If they cannot belive their ears, they can belive their eyes! But even though such expressions are common, people generally underestimate the actual amount of information that is transmitted through the optic channel. Humans produce speech through the actions of several articulators (vocal cords, velum, tongue, lips, jaw, etc.), of which only some are visible. The continuous speech thus produced is not, however, continously audible: It is also made of significant parts of silence, during voiceless plosives and during pauses, while the speaker makes gestures in order to anticipate the following sound. To sum up, parts of speech movements are *only visible*, parts are *only audible*, and part are *not only audible*, but *also visible*. Humans take advantage of the bimodality of speech; from the same source, information is *simultaneously* transmitted through two channels (the acoustic and the optic flow), and the outputs are integrated by the perceiver.

In the folowing discussion, I first pinpoint the importance of visual intelligibility of speech for normal headers, and discuss some of the most recent issues in the bimodal aspects of speech production and perception. Then, I detail various aspects of an emerging technology, since many applications can be based on animated talking faces; very high quality synthetic "actors" can now be animated in real time in the movie industry, and Text-to-Audio-Visual-Speech (TtAVS) synthesizers are needed in the multimodal use of man-machine dialogues.

2. WHAT IS KNOWN ABOUT NATURAL SPEECH

2.1 Intelligibility of Visible Speech

It is well known that lip-reading is necessary in order for the hearing impaired to (partially) understand speech, specifically by using the information recoverable from visual speech. But as early as 1935, Cotton stated that "there is important element of visual hearing *in all normal individuals*" [23] [emphasis mine]. Even if the auditory modality is the most important for speech perception by normal hearers, the visual modality may allow subjects to better understand speech. Note that visual information, provided by movements of lips, chin, teeth, cheeks, etc., cannot, in itself, provide normal speech intelligibility. However, a view of the talker's face enhances spectral information that is distored by background noise. A number of investigators have studied this effect of noise distortion on speech intelligibility according to whether the message is heard only, or heard with the speaker's face also provided [84], [61], [12], [27], [28], [85], etc.

Fig. 1 reports articulation scores obtained in French by Mohamadi and Benoît [57] on 18 nonsense words by 18 normal hearers in two test conditions: audition only and audition plus vision. We observe that vision is basically unnecessary in rather clear acoustic conditions ($S/N > 0\text{dB}$), whereas seeing the speaker's face allows the listener to understand around 12 items out of 18 under highly degraded acoustic conditions ($S/N = -24\text{dB}$) where the auditory alone message is not understood at all.

* This article will appear in *The Structure of Multimodal Dialogue*, 2, D. Bouwhuis, F. Néel, M. Taylor, Eds.

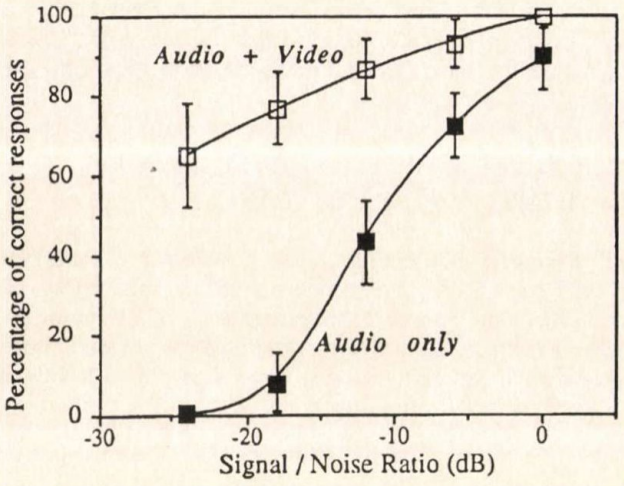


Fig. 1. Improved intelligibility of degraded speech through vision of the speaker's face. The box indicates the mean, and the whiskers the standard deviation.

One may reply that such conditions are seldom found in our everyday lives, only occuring in very noisy environment such as discotheques, in some streets or industrial plants. (Quite fortunately!) But using visual speech is not merely a matter of increasing acoustic intelligibility for hearers/viwers: it is also a matter of making it more comprehensible, i.e., easier to understand. It is well known that information is more easily retained by an audience when transmitted over the television than over the radio. To confirm this, Reisenberg et al. [75] reported that passages read from Kant's *Critique of Pure Reason* were better understood by listeners (according to the proportion of correctly repeated words in a shadowing task) when the speaker's face was provided to them. Even if people usually do not speak the same way as Emmanuel Kant wrote, this last finding is clear argument of the general overall improvement of linguistic comprehension through vision. Therefore, it also allows us to better take into consideration the advantage of TtAVS synthesis for the understanding of automatically read messages, assuming that human-machine dialogue will be much more efficient under bimodal presentation of spoken information to the user.

An avarage 11 dB "benefit of lip-reading" was found by MacLeod and Summerfield [44]. This corresponds to the average difference between the lowest signal-to-noise ratios at which test sentences are understood, given presence or absence of visual information. This finding must obviously be tempered by the conditions of visual representation. Östberg et al. [63] tested the effects of six sizes of videophone display on the intelligibility of noisy speech. They presented running speech to subjects who were asked to adjust the noise level so that the individual words in the story appeared at the borderline of being intelligible; they observed an increase in the mean benefit of lip-reading from 0.4 to 1.8 dB with the increase in the display size. This observation confirms the intuitive idea that the better the visual information, the greater the improvement in intelligibility.

2.2 The Need for Coherence between Facial Gestures and Speech Sounds

The main problem researchers have to deal with in the area of speech production and bimodal speech perception (by ear and by eye) is the coherence of the acoustic and the visual signals (see [26] [47] [19], for extended discussions of this phenomenon). I will briefly present experimental results obtained from perceptual studies where various kinds of coherence were not respected: When the auditory and visual information channels have spatial, temporal, or source differences.

2.2.1 Spatial Coherence

It has been established that either modality influences spatial localization of the source through the other [11]: Subjects who are instructed to point at a visual source of information deviate slightly from it if a completing acoustic source is heard from another spatial position, and conversely, subjects deviate more from the original acoustic source if a competing optical source interferes from another location. In speech, such a "capture of the source" is well known and widely used by ventriloquists, as the audience is much more attracted by the dummy whose facial gestures are more coherent with what they hear than those of its animator [90]! Even four-to-five month old infants, presented simultaneously with two screens displaying video films of the same human face, are preferentially attracted by a face pronouncing the sounds heard rather than a face pronouncing something else [39]. This demonstrates a very early capacity of human to identify coherence in the gestures and their corresponding acoustic production. This capacity is frequently used by listeners in order to improve the intelligibility of a single person in a conversation group, when the well-known "cocktail party effect" occurs.

2.2.2 Temporal Coherence

The second problem which arises from the bimodal aspect of speech perception is due to the inherent synchrony between acoustically and optically transmitted information. Dixon and Spitz [24] have experimentally observed that subjects were unable to detect asynchrony between visual and auditory presentation of speech when the acoustic signal was presented less than 130 ms before or 260 ms after the continous video display of the speaker's face. (Note that this delay sensitivity is much more accurate in case of a punctual event, such as a hammer hitting an anvail, where the range is from 75 ms before to 190 ms after.) Mainly motivated by the applied problem of speech perception through the visiophone (where the unavoidable image coding/decoding process delays the transmission of optical information), recent studies tried to quantify the loss of intelligibility due to delayed visual information. For example, Smeele and Sitting [82] measured the intelligibility of phonetically balanced lists of nonsense CVC words acoustically degraded by background interfering prose. They measured a mean intelligibility of 20% in

the auditory alone condition of presentation and of 65% in the audio-visual condition. However, if the facial presentation was delayed more than 160 ms after the corresponding audio signal, there was no significant improvement of audio-visual presentation over audio alone. In the other direction, Smeele (personal communication) more recently observed a rather constant intelligibility of around 40% when speech was presented in a range of 320 to 1500 ms after vision. In a similar experiment, Campbell and Dodd [17] had previously discovered that the disambiguisation effects of speech-reading on noisy isolated words were observed with durations of up to 1.5 sec desinchrony between seen and heard speech, but they indicated that this benefit occured whichever modality was leading. On the other hand, Reisenberg et al. [75] failed to observe any visual benefit in a shadowing task using the above mentioned text by Kant with modalities desynchronised at 500 ms. These – somewhat divergent – findings strongly support the idea that audition and vision influence each other in speech perception, even if the extent of the phenomenon is unclear (i.e., does it operate on the acoustic feature, the phoneme, the word, the sentence, etc.?) and even if the role of short-term memory, auditory and visual, in their integration remains a mystery.

I would simply suggest that the benefit of speech-reading is a function not only of the acoustic degradation, but also of the linguistic comlexity of the speech material. The grater the redundancy (from nonsense words to running speech through isolated words), the more the high-level linguistic competence is solicited (in order to take advantage of the lexicon, syntax, semantics, etc., in a top-down process), and the more this cognitive strategy dominates the low-level bottom-up decoding process of speech-reading.

2.2.3 Source Coherence

Roughly speaking, the phoneme realizations that are most easily discriminable by the ears are those which are the most difficult to distinguish by the eyes, and vice versa. For instance, /p/, /b/, and /m/ look alike in many languages, although they obviously sound unlike, and are often grouped together as one *viseme*. On the other hand, speech recognizers often make confusions between /p/ and /k/, whereas they look very different on the speaker's lips. This implies that a synthetic face can easily improve the intelligibility of speech synthesizer – or of a character'c voice in a cartoon – if the facial movements are *coherent* with the acoustic flow that is supposed to be produced by them. If not, any *contradictory* information processed during the bimodal integration by the viewer/listener may greatly damage the intelligibility of the original message. This dramatic effect can unfortunately result if the movements of the visible articulators are driven by the acoustic flow, e.g., through an acoustic-phonetic decoder. Such a device might involuntarily replicate the well-known *McGurk effect* [54], where the simultaneous presentation of an acoustic /ba/ and of a visual /ga/ (a predictable decoder error) makes the viewer/listener perceive a /da/! I must emphasize that the McGurk effect is very compelling, as subjects who are well aware of the nature stimuli even fall for the illusion. Moreover, Green et al. [33] found little difference in the magnitude of the McGurk effect between subjects for whom the sex of the voice and the face presented were either matched or mismatched. They concluded that the mechanism for integrating speech information from the two modalities is insensitive to certain incompatibilities, even when they are perceptually apparent.

2.3 The Specific Nature of Speech Coherence between Acoustics and Optics

Speaking is not the process of uttering a sequence of discrete units. Coarticulation systematically occurs in the transitions between the realizations of phonological units. Anticipation or perseveration across phonetic units of articulator gestures in the vocal tract are well known for their acoustic consequences, i.e., for the differences in allophones of a single phoneme. In French, for instance, the /s/, which is considered a non-rounded consonant, is spread in /si/, but proturded in /sy/, due to regressive assimilation; on the opposite, /i/, which has the phonological status of a "spread" phoneme, is protruded in /i/, due to progressive assimilation (which is less frequent). Such differences in the nature of allophones of the same phonemes are auditorily pertinent, and visually pertinent [18], [21].

A classic example of anticipation in lip rounding was first given by Benguérel and Cowan [6] who observed an articulatory influence of the /y/ on the first /s/ in /istrstry/ which occured in the French sequence *une sinistre structure* (though this has since been revised in [2]). In fact, the longest effect of anticipation is observed during pauses, when no acoustic cues are provided, so that subjects are able to visually identify /y/ an average of 185 ms before it is pronounced, during a 460 ms pause in /i, y/ [20].

In an experiment where subjects had simply to define the final vowel in /zizi/ or /zizy/, Escudier at al. [30] showed that subjects visually identified the /y/ in /zizy/ from a photo of the speaker's face taken at around 80 ms before the time when they were able to auditorily identify it (from gated experts of various lengths of the general form /ziz.../)/ They also observed no difference in the time when subjects could identify /i/ or/y/, auditorily or visually, in the transitions /zyzi/ or /zyzy/. This asymmetric phenomenon is due to non-linearities between articulatory gestures and their acoustic consecuences [83]. In this example, French speakers can round their lips – and they do so! –before the end of the /i/ in /zizy/ without acoustic consequences, whereas spreading the /y/ too early in /zyzi/ would lead to a mispronounciation and therefore to a misidentification. To acoustically produce a French /y/, lips have to be rounded so that their interolabial area is less than 0,8 cm^2, above which value it is perceived as /i/ [1]. Lip control is therefore much more constrained for /y/ than for /i/, leading to an anticipation of lip rounding in /i/ /y/ transitions longer than that of lip spreading in /y/ /i/ transitions.

We see from these observations that coarticulation plays a great role in the possibilities for subjects to process visual information before, or in abscence of, acoustic information. This natural asynchrony between the two modes of speech perception depends upon the intrinsic nature of phonetic units, as well as on the speech rate and individual strategy of the speaker. It is obvious that the increase of intelligibility given by vision to audition relies on it.

2.4 Bimodality of Speech

2.4.1 Synergetic Bimodality of Speech

1% + 6% = 45%! Setting communications parameters at threshold level, Risberg and Lubker observed that when a speaker appeared on a video display, but with the sound turned off, subjects relying on speech-reading correctly perceived 1% of test words [76]. When the subjects could not see the display, but where presented with a low-pass filtered version of the speech sound, they got 6% correct. Presented with the combined information channels, the performance jumped to 45% correctly perceived test words. This observation exemplifies the remarkable synergy of the two modes of speech perception. However, little is known about the process that integrates the cues across modalities, although a variety of approaches and of models to multimodal integration of speech perception have been proposed [47], [86], [14], and tested [50], [52], [32], [49], [77]. Our understanding of this process is still relatively crude, but its study is very active and controversial at present (see the 21 remarks to, and in, [48]!).

2.4.2 Specific Bimodality of Speech

To sum up these various psycholinguistic findings: as concerning speaker localization, vision is dominant on audition; for speaker comprehension, vision greatly improves intelligibility, especially when acoustics is degraded and/or the message is complex; this speech-reading benefit generally holds even when the channels are slightly desynchronized; due to articulatory anticipation, the eye often receives information before the ear, and seems to take advantage of it; and finally, as for localization, vision can bias auditory comprehension, as in the McGurk effect.

The Motor Theory of speech perception [42] supposes that we have an innate knowledge of how to produce speech. Recently, in a chapter of a book devoted to the reexamination of this theory, Summerfield [87] suggested that the human ability to lipread could also be innate. His assumption allows a partial explanation of the large variability observed in human performance at speech-reading, as this ability seems to be related to the visual performance capacities of the subject [80] [79]. Summerfield also hypothesized that evolutionary pressure could led to refined authory abilities for biologically significant sounds, but not for lipreading. Therefore, whatever the innate encoding of speech, whether in an auditory or visual form. an intermediate stage of motor command coding allowing us to perceive speech would provide us not only with the coherence of acoustic and visual signals in a common metric, but also with an improvement in the processing of the speech percept (whose final storage pattern is still an open question). This is my interpretation of the famous formula "Perceiving is acting," recently revised by Viviani and Stucchi into "Perceiving is knowing how to act" [89].

3. HOW TO DEAL WITH SYNTHETIC SPEECH

3.1 Annimation of Synthetic Faces

In the last two decades, a variety of synthetic faces have been designed all over the world with the objective of their animation. The quality of "facial models" goes from a simple electronic curve on an oscilloscope, through a wide range of pre-stored human face shapes and more or less caricatural $2D$ vector-driven models of the most salient human face contours, to a very natural rendering of a $3D$ model (by mapping of real photographs on it, for instance).

I think that, as in the case of acoustic speech synthesis, one must differentiate the final model from its animation technique. I suggest the reader refer to Broke for a tutorial presentation of the techniques and methods used in a facial animation [16]. To sum up, Table 1 gives a classification of the most noticeable publications of designed systems along these two criteria. For simplification purposes, Table 1 does not consider the work done by investigators to develop or apply the models cited, nor to synchronise them with speech synthesizers. Of course, the control parameters of a rule-driven model may be given by a code-book, after analysis of the acoustic wave-form, so the Y-axis could have been presented differently. However, the table aims at showing the basic characterictics of the various approaches by assigning the most suitable animation technique to a given model.

Table 1. Classification of best-known face synthesizers based on the kind of facial model developed (X-axis), and on its underlying method of animation (Y-axis). Only the first author an date of publication are quoted; complete references are given at the and of the article.

Facial model: / Animation technique:	Lissajou-like (Electronics)	2D vectors or shapes	3D wire or raster-graphics
from acoustics	Boston, 1973 Erber, 1978	Simons, 1990 (through stochastic networks)	Morishima, 1990
by rules (control parameters)		Brooke, 1979	Parke, 1974 Platt, 1981 Waters, 1987 Magnenat-Thalmann, 1988
by code-books (quantization) or key-frames (interpolation)		Montgomery, 1982 Matsuoka, 1988 Woodward, 1991 Mohamadi, 1992	Parke, 1972 Bergeron, 1985 Aizawa, 1987 Nahas, 1988

Whatever the facial model, it may be animated for speech by three main methods:

- A direct mapping from acoustics to geometry creates a correspondance between the energy in a filtered bandwidth of the speech signal and the voltage input of an oscilloscope, and therefore with the horizontal or vertical size of an elliptical image [13], [29]. An indirect mapping may also be achieved through a stochastic network which outputs a given image from a parametrization of the speech signal, after training of the network in order to make it match any possible input with its corresponding output as accurately as possible. Simons and Cox [81] used a Hidden Markov network, whereas Morishima et al. [59] preferred a connexionist network, for instance. In both, the inputs of the networks were LPC parameters calculated from the speech signal.
- When the facial model has previously been designed so that it can be animated through control parameters, it is possible to elaborate rules which simulate the gestural movements of the face. These commands can be based on a geometrical parametrization, such as jaw rotation or mouth width for instance, in the case of the Brooke's $2D$ [15] or Parke's $3D$ model [67], or on an anatomical description, such as the muscle actions simulated in the $3D$ models by Platt and Badler [73], Waters [91], or Magnenat-Thalmann et al [45].
- Facial models may only mimic a closed set of human expressions, whatever the tool used to create them: a set of $2D$ vectors [58], [94]; simplified photos [56]; hand-drawn mouth-shapes [53]; $3D$ recontructions of $2D$ images [66]; $3D$ digitizing of a mannequin using a laser scanner [46], [60], [38], [40]; or even direct computer-assisted sculpting [64]. If no control parameters can be applied to the structure obtained in order to deform it and so generate different expressions, and if digitizing of multiple expressions is possible, hand-modification by an expert is necessary in order to create a set of relevant key-frame images. The pre-stored images can then be concatenated as in cartoons, so that a skilled animator may achieve coherent animation. Such a technique has been widely employed, since it only requires a superficial model of face, and as little physiological knowledge of speech production is needed to modify the external texture so that natural images can be duplicated (*rotoscoping technique*).

I also want to mention two other techniques that directly rely on human gestures:

- "Expression slaving" consits of the automatic measurement of geometrical characteristics (reference point or anatomical measurements) of a speaker's face, which are mapped onto a facial model for local deformation [10], [88], [93], [68]. In the computed-generated animation *The Audition* [55], the facial expressions of a synthetic dog are mapped onto its facial model from natural deformations automatically extracted from a human speaker's face [68].
- Attempts have also been made to map expressions to a facial model from control commands driven in real-time by the hand of a skilled puppeteer [25]. In a famous computer-generated French TV series, *Canaille Peluche*,* the body and face gestures of Mat, the synthetic ghost, are so created. Finally, a deformation that is rather based on the texture than on the shape of the model can be achieved by mapping of various natural photos on the model [40].

The basic principles of these various animation techniques are represented in Fig. 2.

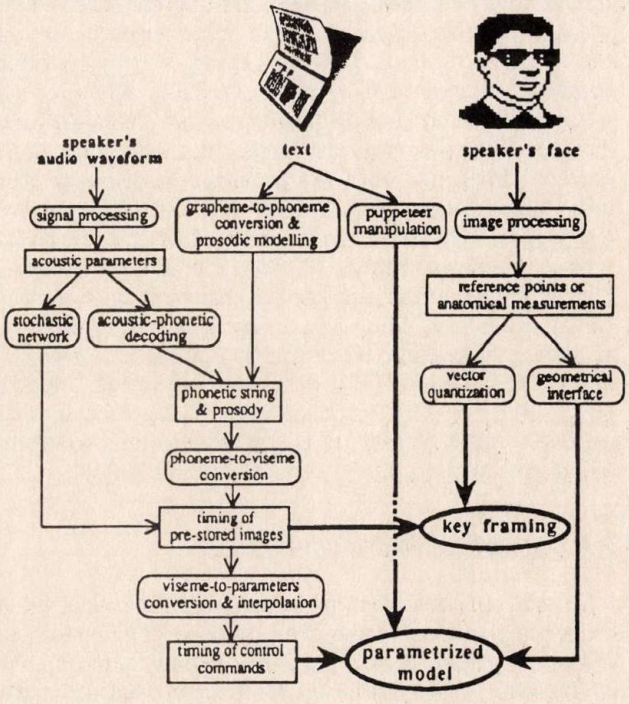


Fig 2. General scheme of Facial Animation showing the possible techniques.

3.2 Audio-Visual Speech Synhesis

Experiments on natural speech (see Section 2.2 above) allow us to anticipate that similar effects will be obtained with a TtAVS synthesizer: Even if the current quality of (most) TtS systems is not as bad as highly degraded speech, it is obvious that under very quiet conditions, synthesizers are much less intelligible than humans. Moreover, it is realistic to predict than in the naer future, the spread of speech syntesizers will lead to wide use in noisy backgrounds, such as in railway stations. Such adverse conditions will necessitate a synchronized presentation of the information from another modality, for instance, the orthographic display of the text, or the animation of a synthetic face (especially for foreigners and illitirates). There are hence several reasons for the study and use of Audio-Visual Speech Synthesis.

Audio-Visual Synthesis allows to accurately control stimuli for perceptual tests on bimodality: Massaro and Cohen [51] studied how speech perception is influenced by information presented to ear and eye by dubbing acoustiy tokens generated by a speech synthesizer [36] onto a sequence of images generated by a facial model [67], [69]. Highly controlled synthetic stimuli thus allowed to investigate in details the McGurk effect. Audio-Visual Synthesis is also a tool for basic research on speech production: Pelachaud [70] studied the relationship between intonation and facial expressions by means of natural speech and the facial model developed by Platt [72].

Thanks to the increasing capacities of computer graphics, highly natural (or hyper-realistic) rendering of $3D$ syn-

* Produced by Canal Plus and Videosystem

thetic faces now allows movie producers to create synthetic actors whose facial gestures have to be coherent with their acoustic production, due to their human-like quality and the demands of the audience. Short computer-generated movies clearly show this new trend: *Tony de Peltrie* [10]; *Rendez-vous à Montréal* [46]; *Sextone for president* [37]; *Tin Toy* [74]; *Bureaucrat* [92]; *Hi Fi Mike* [96]; and *Don Quichotte* [31], among others. It is necessary for computer-assisted artists to be equipped with software facilities so that the facial gestures and expressions of their characters are easily, quickly, and automatically generated in a coherent manner.

Several attempts to synchronize synthetic faces with acoustic (natural or synthetic) speech may be found in literature: [41], [35], [53], [60], [22], [59], [71], [95]; among others, for (British & American) English, Japanese, and French. Unfortunately, most of the autors only reported informal impressions from collegueas about the quality of their system, but – as far as I am aware – none of them has ever quantified the improvement in intelligibility given by adding visual synthesis to the acoustic waveflow. I strongly support the idea that assessment methodologies should be standardized so that the various approaches can be compared to one another. As proposed by Benoît & Pols [9], a method similar to that reported above in section 2.1 should be systematically carried out by designers. As regards my own culpability, I note that such an experiment is under way at ICP, and results are to be published by the end of the year...

3.3 The TtAVS System under Development at ICP

Prior to detailing the method that was adopted for the design of a Text-to-Audio-Visual-Speech synthesizer in our laboratory, I will detail its characteristics.

3.3.1 The Control of the Coarticulation Effect

Coarticulation is the most difficult problem investigators have to deal with in the animation of talking faces, partially because it is speech-specific and language-dependent. Coarticulation involves the effect of facial expressions (such as a smile, for instance) on the gestures of the mouth and jaw, and the assimilation effect of the phonetic context on the production of a sound. Little work has been undertaken thus far in order to control the visible effects of coarticulation in visual speech synthesis, with the notable exception of a dissertation recently devoted to this important problem [70]. To give two examples of the coarticulation effect, it has been observed by Benoît et al. that the general mouth/jaw shape of a French speaker is the same when he utters a steady /*i*/ or when he utters an /*a*/ in the /$\zeta a \zeta$/ sequence, and that the /*i*/ in a /$\zeta i \zeta$/ sequence is of the same shape as the surrounding /ζ/.

To take this into account, Benoît at al. [8] measured a set of relevant geometrical and anatomical parameters on the face of a speaker uttering numerous combinations of French coarticulated vowels and consonants. Multidimensional data analysis allowed identification of a set of around twenty key shapes (termed *visemes*) that are meant to structurally describe the articulatory space covered by the visible gestures of a Frenchman speaking with natural expressions. Such a quantitive description of the facial movements involved in the production of coarticulated speech serves as a theoretical basis for the definition of the key frames used in our TtAVS system.

3.3.2. The Choice of a Key-Frame-Based System

The most simple way to design a prototypical face synthesizer is to develop a key-frame-based system, as no underlying model of the face is needed. The main problem is obviously the selection of relevant images. Indeed, Parish at al. [65] demonstrated that an "intelligent" temporal subsampling of a sequence of photos of a signer uttering a word in sign language was more intelligible than a random subsampling equal in number of photos: the selected photos were those differing with their neighbours, i.e., the most steady, and consequently, the most informative according to the intelligibility of linguistic gestures. This effect is well known and widely used by cartoon animators. As stated above in section 3.1, several designers used a similar facial animation technique.

Saintourens and his colleauges were the first investigators to develop a French TtAVS synthesizer [78]. They synchronized the visual display of a limited set of facial images with the corresponding phonemes being synthesized by the Text-to-Speech system developed at the CNET [34]. To do so, they pre-stored 24 images in the memory of a special graphics board in order to make them accesible in real-time from a PC. These images were previously calculated and rendered on a computer after modification of a 3*D* scanned face and application of B-spline functions, so that the main expressions that can occur during (netural) speech could be represented. The result is of high visual quality, but it does not take into account coarticulation, leading to occasional unnatural movements of the face.

Using a similar principle, a TtAVS synthesizer is being developed at the ICP. It allows real time display on a PC VGA screen of around 25 key frame images per second out of a set of about 23 images, being either graphic contours of a vector-drawn face [94], or binary photos of the lower part of the front view of a speaker's face [56]. In both cases, each image corresponds to one of the visemes previously identified [8]. Such a low-cost TtAVS system is currently implemented at the ICP [95], using the facilities of the COMPOST rule compiler [5] through which the acoustic synthesis of speech from unrestricted French text was previously developed [4].

As shown in Fig.3, there are three main steps in the structure of phoneme-to-synthetic face conversion, all using rewriting rules between two limited vocabulaires at the symbolic level. The first consists of a simple matching between each phoneme and its corresponding *archi-viseme*; e.g., /p, b, m/, are assigned the same [B] output, whatever their phonetic context. The second step consists of rewriting these archi-visemes into visemes, under contextual constraints; e.g., [A] preceded and followed by [Z] is rewritten (I) (see example above). The graphic database contains a facial image corresponding to each of these visemes. Then, in the final step, a set of rewriting rules aligns the images with the acoustic output, doubling an image (or inserting a different image if an appropriate interpolated image exists) in between any two others as the timing requires. An

ad hoc program allows the visual display and the acoustic waveform so generated to be synchronized, either in real time, or after storage of the acoustic signal in one file and the sequence of the image codes and their timing in another.

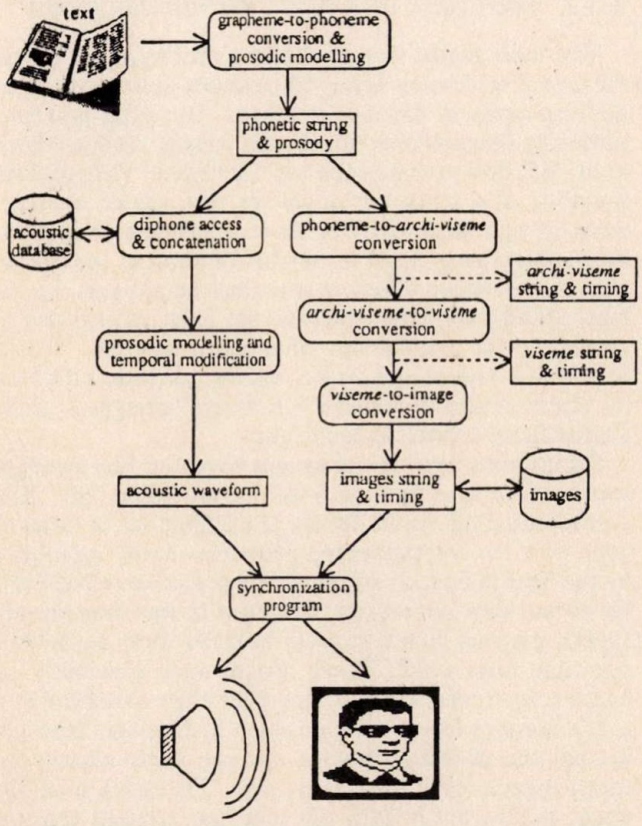


Fig. 3. Simplified principle of the ICP Text-to-Audio-Visual-Speech synthesizer

ASCII files of all intermediate outputs are created, including the duration of each symbolic unit, so that this techniques eases manual modifications of any rule, as well as any hand-labeling of natural speech for post-synchronization of natural audio and synthetic video signals.

In the near future, the assessment of the intelligibility of this system will be carried out with the methodology presented in section 2.1. This will lead to an evaluation of the intelligibility contributions of synthetic voice and synthetic face under various conditions of acoustic degradation, as compared with natural speech.

4. PERSPECTIVES IN THE SYNTHESIS OF BIMODAL SPEECH

A much higher quality of synthetic face planned at the ICP by means of an automatically parametrized $3D$ facial model (driven from an ortographic input or from an analysis of the natural gestures of speaker) and synchronized with the acoustic output. The same procedure as stated above will thus be used in order to qualify the intelligibility of such a talking face.

To end in the vein of Negroponte's dream quoted at the begining of this article, I think that an idealistic TtAVS synthesizer, or a holographically-rendered intelligent speaking robot will be constituted of a centralized parametric model of the vocal tract, such as that developed by Maeda [43], which would simultaneously synthesize the acoustic waveform and control the visible part of the vocal tract in a facial model. This will be the ultimate case where both modalities of the speech are *coherent*, since they are produced by the same source, i.e., the same anthropomorphic model.

Prior to reaching this goal, the developement of basic research in this fascinating area urgently needs the set-up of a multilingual collaborative project that would aim at storing multimedia databases where natural speech would be synchronously labelled on the acoustic waveform and on the front and profile views of the speakers. Europe should see to this!

ACKNOWLEDGEMENTS

I am most indebted to Christian Abry, Louis-Jean Boë, and Marie-Agnès Cathiard for patience, comments, and help, especially in the bibliographical labyrinth, and in my attempts to order it. As in many other occasions, I am deeply grateful to the Scientific Editor who best knows how translating Benoît's Frenchglish into American English, namely Tom Sawallis.

REFERENCES

[1] Abry, C., Boë, L. J., and Schwartz, J. L., "Plateaus, catastrophes and structuring of vowel systems", *Journal of Phocetics*, vol. 17, 1989, 47–54.

[2] Abry, C., and Lallouache, M. T., "Audibility and stability movements: Deciphering two experiments on anticipatory rounding in French", *Proceedings of the XIIth Intenational Congress of Phonetic Sciences*, Aix-en-Provence, France, vol. 1, 1991, pp. 220–225.

[3] Aizawa, K., Harashima, H., and Saito, T., "Model-based synthetic image coding", *Proceedings of Picture Coding Symposium*, vol. 3 (11), 1987, pp. 50–51.

[4] Bailly, G., and Guerti, M., "Synthesis-by-rule for French", *Proceedings of the XIIth International Congress of Phonetic Sciences*, Aix-en-Provance, France, vol. 2, 1991, pp. 506–511.

[5] Bailly, G., and Tran, A., "Compost: A rule-compiler for speech synthesis", *Proceedings of the Eurospeech Conference, Paris, France*, 1989, pp. 136–139.

[6] Benguérel, A. P., and Cowan, H. A., "Coarticulation of upper lip protrusion in French", *Phonetica*, vol. 30, 1974, pp.41–55.

[7] Benoît, C., Boë, L. J., and Abry, C., "The effect of context on labiality in French", *Proceedings of the 2nd Eurospeech Conference*, Genova, Italy, 1991, pp. 153–156.

[8] Benoît, C., Lallouache, M. T., Mohamadi, T., and Abry, C., "A set of French visemes for visual speech synthesis", in *Talking Machines: Theories, Models and Designs, G. Bailly and C. Benoît, Eds, Elsevier Science Publishers B. V.*, North-Holland, Amsterdam, 1992, pp. 485–504.

[9] Benoît, C., and Pols, L. C. W., "On the assessment of synthetic speech", in *Talking Machines: Theories, Models and Designs, G. Bailly and C. Benoît, Eds, Elsevier Science Publishers B. V.*, North-Holland, Amsterdam, 1992, pp. 435–442.

[10] Bergeon, P., and Lachapelle, P., "Controlling facial expressions and body movements in the computer generated animated short 'Tony de Pelterie'", *SigGraph '85 Tutorial Notes*, Advanced Computer Animation Course.

[11] Bertelson, P., and Radeau, M., "Cross-modal bias and perceptual fusion with auditory visual spatial discordance", *Perception and psychophysics*, vol. 29, 1981, pp. 578–584.

[12] innie, C. A., Montgomery, A. A., and Jackson, P. L., "Auditory and visual contributions to the perception of consonants", *Journal of Speech & Hearing Research*, vol. 17, 1974, pp.619–630.

[13] Boston, D. W., "Synthetic facial animation", *British Journal of Audiology*, vol. 7, 1974, pp. 373–378.

[14] Broida, L. D., "Crossmodal integration in the identification of consonant segments", *Quaterly Journal of Experimental Psychology*, vol. 43, 1991, pp. 647–678.

[15] Brooke, N. M., "Development of a video speech synthesizer", *Proceedings of the British Institute of Acoustics*, Autumn Conference, 1979, pp. 41–44.

[16] Brooke, N. M., "Computer graphics synthesis of talking faces", in *Talking Machines: Theories, Models and Designs, G. Bailly and C. Benoît, Eds, Elsevier Science Publishers B. V.*, North-Holland, Amsterdam, 1992, pp. 505–522.

[17] Campbell, R, and Dodd, B., "Hearing by eye", *Quaterly Journal of Experimental Psychology*, vol. 32, 1980, pp. 509–515.

[18] Cathiard, M. A., "Identification visuelle des voyelles et des consomes dans le jeu de la protrusion-rétraction des lèvres en français", *Mémoire de Maîtrise*, Département de Psychologie, Grenoble, France, 1988.

[19] Cathiard, M. A., "La perception visuelle de la parole: aperçu des connaissances", *Bulletin de l'Institut de Phonétique de Grenoble*, vol. 17/18, 1988/1989, pp. 109–193.

[20] Cathiard, M. A., Tiberghien, G., Tseva, A., Lallouache, M. T., and Escudier, P., "Visual perception of anticipatory rounding during pauses: A cross-langue study", *Proceedings of the XIIth International Congress of Phonetic Sciences*, Aix-en-Provance, France, vol. 4, 1991, pp. 50–53.

[21] Chafcouloff, M., and Di Cristo, A., "Les indices acoustiques et perceptuels des consonnes constrictives du français, application à la synthèse", *Actes des 9èmes Journées d'Etude sur la Parole*, Grouppe Communication Parlée du GALF, Lannion, France, 1978, pp. 69–81.

[22] Cohen, M. M., and Massaro, D. W., "Synthesis of visible speech", *Behaviour Research Methods, Instruments & Computers*, vol. 22(2), 1990, pp. 260–263.

[23] Cotton, J., "Normal 'visual-hearing'", *Science*, vol. 82, 1935, pp. 592–593.

[24] Dixon, N. F., and Spitz, L., "The detection of audiovisual desynchrony", *Perception*, vol. 9, 1980, pp. 719–721.

[25] deGraf, B., "Performance facial animation notes", *Course Notes on State of the Art in Facial Animation*, SigGraph '90, vol. 26, 1990, pp. 10–20.

[26] Dodd, B., and Campbell, R. (Eds), "Hearing by eye: The Psychology of lip-reading", Lawrence Erlbaum Associates, Hillsdale, New Jersey 1987.

[27] Erber, N. P., "Interaction of audition and vision in the recognition of oral speech stimuli", *Journal of speech & Hearing Research*, vol. 12, 1969, pp.423–425.

[28] Erber, N. P., "Auditory-visual perception of speech", *Journal of speech & Hearing Disorders*, vol. 40, 1975, pp.481–492.

[29] Erber, N. P., and De Filippo, C. L., "Voice/mouth synthesis and tactual/visual perception of /pa, ba, ma/", *Journal of the Acoustical Society Of America*, vol. 64, 1978, pp. 1015–1019.

[30] Escudier, P., Benoît, C., and Lallouache, M. T., "Identification visuelle de stimuli associés à l'opposition /i/–/y/: étude sttique", *Proceedings of the First Conference on Acoustics, Lyon, France*, 1990, pp.541–544.

[31] Garnier, F., *Don Quichotte*, Computer-generated movie, Videosystem, Paris, France A. Guidot Prod., 2' 40" (1991).

[32] Grant, K. W., and Braida, L. D., "Evaluating the articulation index for auditory-visualinput", *Journal of the Acoustical Society of America*, vol. 89, 1991, pp. 2952–2960.

[33] Green, K. P., Stevens, E. B., Kuhl, P. K., and Meltzoff, A. M., "Exploring the basis of the McGurk effect: Can perceivers combine information from a female face and a male voice?", *Journal of the Acoustical Society of America*, vol. 87, 1991, pp. S125.

[34] Hamon, C., Moulines, E., and Charpentier, F., "A diphone synthesis system bases on time-domain prosodic modifications of speech", *Proceedings of the IEEE Conference on Acoustics Speech & Signal Processing*, 1989, pp. 1989.

[35] Hill, D. R., Pearce, A., and Wyvill, B. L. M., "Animating speech: an automated approech using speech synthesized by rules", *The Visual Computer*, vol. 3, 1989, pp. 277–289.

[36] Klatt, D. H., "Software for a cascade/parallel format synthesizer", *Journal of the Acoustical Society of America*, vol. 67, 1980, pp. 971–995.

[37] Kleiser, J., *Sextone for President*, Computer-generated movie, Kleiser-Walczak construction Comp., 28" (1988).

[38] Kleiser, J., "A fast efficient accurate way to represent the human face", *Course Notes on State of the art in Facial Animation*, SigGraph '89, vol. 22, pp.35–40.

[39] Kuhl, P. K., and Meltzoff, A. N., "The bimodal perception of speech in infancy", *Science*, vol. 218, 1982, pp. 1138–1141.

[40] Kurihara, T., and Arai, K., "A transformation method for modelling and animation of the human face from photographs", *Computer Animation '91*, N. Magnenat-Thalmann & D. Thallmann, Eds, Springer-Verlag, 1991, pp. 45–58.

[41] Lewis, J. P., and Parke, F. I., "Automated lip-synch and speech synthesis for character animation", *Proceedings of CHI '87 and Graphics Interface '87,*, Toronto, Cannada, 1987, pp.143–147.

[42] Liberman, A., and Mattingly, I., "The Motor Theory of Speech Perception Revisited", *Cognition*, vol. 21, 1985, pp. 1–33.

[43] Maeda, S., "Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal tract shapes using an articularory model", in *Speech Production an Speech Modeling*, W. J. Hardcastle & A. Marchal, Eds, Kluwer Academic pubs, 1991, pp. 131–149.

[44] MacLeod, A., and Summerfield, Q., "Quantifying the contribution of vision to speech perception in noise", *British Journal of Audiology*, vol. 21, 1987, pp. 131–141.

[45] Magnenat-Thalmann, N., Primeau, E., and Thalmann, D., "Abstract muscle action procedures for human face animation", *Visual Computer*, vol. 3, 1988, pp. 290–297.

[46] Magnenat-Thalmann, N., and Thalmann, D., "The direction of synthetic actors in the film Rendez-Vous a Montréal", *IEEE Computer Graphics & Applications*, vol. 7(12), 1987, pp. 9–19.

[47] Massaro, D. W., *Speech perception by ear and eye: a paradigm for psychological inquiry*, Lawrence Erlbaum Associates, Hillsade, New Jersey (1987).

[48] Massaro, D. W., "Multiple book review of 'speech perception by ear and eye...'", *Behavioral and Brain Sciences*, vol. 12, 1989, pp. 741–794.

[49] Massaro, D. W., "Connexionist models of speech perception", *Proceedings of the XIIth International Congress of Phonetic Sciences*, Aix-en-Provance, France, vol. 2, 1991, pp. 94–97.

[50] Massaro, D. W., and Cohen, M. M., "Evaluation and integration of visual and auditory information in speech perception", *Journal of Experimental Psychology: Human Perception & Performance*, vol. 9, 1983, pp.753–771.

[51] Massaro, D. W., and Cohen, M. M., "Perception of synthesized audible and visible speech", *Psychological Science*, vol. 1, 1990, pp. 55–63.

[52] Massaro, D. W., and Friedman, D., "Models of integration given multiple sources of information", *Psychological Review*, vol. 97, 1990, pp. 225–252.

[53] Matsuoka, K., Masuda, K., and Kurosu, K., "Speechreading trainer for hearing-impaired children", in *Training Human Decision Making and Control*, J. Patric & K. D. Duncan, Eds, Elsevier Science Publishers B. V., North-Holland, Amsterdam 1988, pp.153–162.

[54] McGurk, H., and MacDonald, J., "Hearing Lips and Voices", *Nature*, vol. 264, 1976, pp.746–748.

[55] Miller, G. S. P., *The Audition*, Computer-generated movie, Apple Computer Inc., Cupertino, USA, 1990, 3' 10".
[56] Mohamadi, T., "Contribution à la synthèse de visages parlants", *Thèse de Doctorant Institut National Polytechnique*, Grenoble, France, 1992.
[57] Mohamadi, T., and Benoît, C., "Apport de la vision du locuteur à l'inlelligibilite de la parole bruitée", *Bulletin de la Communication Parlée*, 2, Chaiers de l'ICP, INPG, Grenoble, France, 1992.
[58] Montgomery, A. A., and Soo Hoo, G., "ANIMAT: A set of programs to generate, edit and display sequences of vector-based images", *Behavioral Research Methods and Instrumentation*, vol. 14, 1982, pp. 39—40.
[59] Morishima, S., Aizawa, K., and Harashima, H., "A real-time facial action image synthesis driven by speech and text", *Visual Communication and Image processing '90*, The Society of Photo optical Instrumentation Engineers, vol. 1360, 1990, pp.1151—1158.
[60] Nahas, M., Huitric, H., and Saintourens, M., "Animation of a B-spline figure", *The Visual Computer* vol. 3, 1988, pp. 272—276.
[61] Neely, K. K., "Effect of visual factors on the intelligibility of speech", *Journal of yhe Acoustical Society of America*, vol. 28, 1956, pp. 1275—1277.
[62] Negroponte, N., "From Bezel to Proscenium", *Proceedings of SigGraph '89*, 1989.
[63] Östberg, O., Lindström, B., and Renhäll, P. O., "Contribution to speech intelligibility by different sizes if videophone displays", *Proceedings of the Workshop on Videophone Terminal Design*, CSELT, Torino, Italy, 1988.
[64] Pauri, A, Magnenat-Thalmann, N., and Thalmann, D., "Creating realistic three-dimensional human shape characters for computer-generated films", *Computer Animation '91*, N. Magnenat-Thalmann & D. Thalmann, Eds, Springer-Verlag, 1991, pp. 89—99.
[65] Parish, D. H., Sperling, G., and Landy, M. S., "Intelligent temporal subsampling of American Sign Language using event boundaries", *Journal of Experimental Psychology: Human Perceptions & Performance*, vol. 16, 1990, pp. 282—294.
[66] Parke, F. I., "Computer-generated animation of faces", *Proceedings of ACM National Conference*, vol. 1, 1972, pp. 451—457.
[67] Parke, F. I., "A parametric model for human faces", *Ph Dissertation*, University of Utah, Department of Computer Sciences, 1974.
[68] Patterson, E. C., Litwinowitz, P. C., and Greene, N., "Facial animation by spatial mapping", *Computer Animation '91*, N. Magnenat-Thalmann & D. Thalmann, Eds, Springer-Verlag, 1991, pp. 31—44.
[69] Pearce, A., Wyvill, B., Wyvill, G., and Hill, D, "Speech and expression: A computer solution to face animation", *Graphich Interface '86*, 1986, pp. 136—140.
[70] Pelachaud, C, "Communication and coarticulation in facial animation", *Ph.D. thesis*, University of Pennsylvania, USA 1991, p. 240.
[71] Pelachaud, C., Badler, N., and Steedman, M., "Linguistics issues in facial animation", *Computer Animation '91*, N. Magnenat-Thalmann & D. Thalmann, Eds, Springer-Verlag, 1991, pp. 15—30.
[72] Platt, S. M., "A structural model of the human face", *Ph. D. thesis*, University of pennsylvania, USA 1985, p. 216.
[73] Platt, S. M., and Badler, N. I., "Animating facial expressions", *Computer Graphics*, vol. 15(3), 1981, pp. 245—252.
[74] Reeves, W. T., "Simple and complex facial animation: Case studies", in *Course Notes on State of the Art in Facial Animation*, SigGraph '90, vol. 26, 1990.
[75] Reisberg, D., Mclean, J., and Goldfield, A., "Easy to hear but hard to understand: A lip-reading advantage with intact auditory stimuli", in *Hearing by eye: The psychology of lip-reading*, B. Odd & R. Campbell, Eds Erlbaum Associates, Hillsdale, New Jerset, 1987, pp. 97—114.
[76] Risberg, A., and Lubker, J. L., "Prosody and speechreading", *Speech Transmission Laboratory Quaterly Progress & Status Report*, 4, KTH, Stockholm, Sweden, 1978, pp. 1—16.
[77] Robert, J., *Intégration audition-vision par réseaux de neurones: une étude comparative des modèles d'intégration appliqués à la perception des voyelles*, Rapport de DEA Signa-Image-Parole, ENSER, Grenoble, France, 1991.
[78] Saintourens, M., Tramus, M. H., Huitric, H., and Nahas, M., "Creation of a synthetic face speaking in real time with a synthetic voice", *Proceedings of the ESCA Workshop pn Speech Synthesis*, Autrans, France, 1990, pp. 249—252.
[79] Samar, V. J., and Sims, D. C., "Visual evoked componets related to speechreading and spatial skills in hearing-impaired adults", *Journal of Speech & Hearing Research*, vol. 27, 1984, pp. 162—172.
[80] Shepred, D., "Visual-netural correlete of speechreading ability in normal-hearing adults: reliability", *Journal of Speech and Hearing Research*, vol. 25, 1982, pp. 521—527.
[81] Simons, A. D., and Cox, S. J., "Generation of mouthsapes for synthetic talking head", *Proceedings of the Institute of Acoustics*, Great Britain, vol. 12(10), 1990, pp. 475—482.
[82] Smeele, P. M. T., and Sittig, A. C., "The Contribution of Vision to Speech Perception", *Proceedings of the 13th International Symposium on Human Factors in Telecommunications*, Torino, 1990, p. 525.
[83] Stevens, K. N., "The quantal nature of speech: Evidence from articulatory-acoustic data", in *Human communication: A unified view*, E. E. David Jr & P. B. Denes, Eds, McGraw-Hill, New York, 1972, pp. 51—66.
[84] Sumby, W. H., and Pollack, I., "Visual contribution to speech intelligibility in noise", *Journal; of the Acoustical Society of America*, vol. 26, 1954, pp. 212—215.
[85] Summerfield, Q., "Use of visual information for phonetic perception", *Phonetica*, vol. 36, 1979, pp. 314—331.
[86] Summerfield, Q., "Comprehensive account of audio-visual speech perception", in *Hearing by eye: the psychology of lip-reading*, B. Dodd & R. Campbell, Eds, Lawrence Erlbaum Associates, Hillsdale, New Jersey, 1978.
[87] Summerfield, Q., "Visual Perception of Phonetic Gestures", in *Modularity and the Motor Theory of Speech Perception*, G,. Mattingly & Studdert-Kennedy, Eds, Lawrence Erlbaum Associates, Hillsdale, New Jersey, 1991.
[88] Terzopoulos, D., and Waters, K., "Techniques for realistic facial modelling and animation", *Computer Animation '91*, N. Magnenat-Thalmann & D. Thalmann, Eds, Springer-Verlag, 1991, pp. 59—74.
[89] Viviani, P., and Stucchi, N., "Motor-perceptual interactions", in *Tutorials in Motor Behaviour II*, J. Requin & G. Stelmach, Eds Elsevier Science Puplishers B. V., North-Holland, Amsterdam, 1991.
[90] Warren, D. H., Welch, R. B., and McCarthy, T. J., "The role of visual-auditory 'compellingness' in the ventriloquism effect: implications for transitivity among the spatial senses", *Perception and Psycophisics*, vol. 30, 1981, pp. 557—564.
[91] Waters, K., "A muscle model for animating three-dimensional facial expression", *Proceedings of Computer Graphics*, vol. 21, 1987, pp. 17—24.
[92] Waters, K., *Bureucrat*, Computer-generated movie, Schlumberegr Laboratory for Computer Science, Austin, USA, K. Waters, Prod., 1' 22", 1990.
[93] Williams, L., "Performance Driven facial animation", *Computer Graphics*, vol. 24(3), 1990, pp. 235—242.
[94] Woodward, P., "Le speaker de synthèse", *Unpublished DEA Dissertation*, ENSERG, Institut National Polytechnique de Grenoble, France, 1991.
[95] Woodward, P., Mohamadi, T., Benoît, C., and Bailly, G., "Synthèse 'a partir du texte d'un visage parlant français", *Actes des 19èmes Journées d'Etude sur la Parole*, Grouppe Communication Parlée de la SFA, Bruxelles, 1992.
[96] Wyvill, B. L. M., and Hill, D. R., "Expression control using synthetic speech", *SigGrapg '90 Tutorial Notes*, vol. 26, 1990, pp. 186—212.

REAL TIME DIGITAL SPECTOGRAPH FOR TEACHING DEAF CHILDREN

P. P. BODA and L. OSVÁTH

DEPARTMENT OF TELECOMMUNICATIONS AND TELEMATICS
TECHNICAL UNIVERSITY OF BUDAPEST
H-1111 BUDAPEST, STOCZEK U. 2.

A real time spectograph based on an IBM AT compatible computer equipped with a digital signal processing card is presented. The signal processing card is responsible for signal filtering while the PC displays the filtered signal. We show a possible application of the equipment, a hopefully useful tool for teaching deaf children to speak.

1. INTRODUCTION

Present technology allows us to analyse speech signals in real time. It is convenient to observe a spectogram instantly while speaking, instead of waiting several seconds. This kind of real time system is presented in this paper.

In Section 2 we introduce the basic structure of human speech, and a commonly used equipment in phoniatry, the spectograph is shown. The problem of deafness is also discussed, and a method to teach deaf children is described.

The signal filtering is based on Wave Digital Filters. This filter structure allows us to process the speech in real time, and provides a robust 32-band resolution in the 0–4 kHz speech band. This new kind of filter type is introduced in Section 3, together with the basic theory and application.

In Section 4 we present the equipment which can be used to teach deaf children to speak. The application software and some conclusions are also discussed.

2. HUMAN SPEECH AND DEAFNESS

What do we mean by hearing? It is a process by which sound is received and converted into nerve impulses. Any defect of this process causes unability to understand the speech (and, of course, other sounds from the surrounding world too) by hearing. If nerve impulses are not generated, the person is deaf, namely ear-deaf. Generation and perception of human speech is received on three channels: audially (by hearing), logokinestheticly (by perception of muscle motion) and visually (by sight).

The natural way of learning to speak is imitation after hearing which is the simplest method how a well-hearing child can acquire speech. A deaf child cannot be taught this way. (It must be mentioned here that there are several levels of deafness. Present technology can help with individually adapted hearing-aids but it cannot be applied at all for total deafness.)

Secondly, deaf persons have inexperienced speech organs. For this reason, it is a hard way to learn speaking only by feeling the motion of muscles.

Thirdly, one needs a level of ability for comprehension, imagination, concentration and fast thinking to understand the speech visually. There are general requirements for good lip reading, but this is a tiring process for a deaf person. One needs hard work requiring patience and persistence.

In spite of the above mentioned considerations, deaf children can also speak. Their speech may be not so clean, coloured, rich-toned but enough for every day use.

It can be seen that none of the above mentioned methods can be applied exclusively, they have to be used together, supplementing each other. Someone who cannot hear loses the most useful and natural communication channel, so we must help to gain visuality with making the speech visible.

2.1. Electroacoustical examination of speech

If human speech has to be wholly investigated, three features should be examined: time-, frequency and intensity-structure. None of them can be substitued by the others. These three features together characterize speech [1], [11].

The examination can take place on three levels: firstly on the basic independent elements of speech, called phonemes, secondly on the connection of phonemes, and thirdly on contiuous speech. The phonemes can be divided into two classes. Simple phonemes can be pronounced by themselves which means that their frequency- and intensity-structure do not change significantly during the pronounciation. In Hungarian, these are all the vowels and the m, n, v, f, z, sz, zs, s, h consonants. The rest are the combined phonemes: their structure change during the utterance, so their spectrum is not constant during the pronounciation time. In most cases, this happens because the place of the obstacle changes during the pronouncation. These are p, d, t, g, k, gy, ty, j, cs, c, dz, dzs, ny, r, l.

We obtain the time-structure of speech by considering the frequency and intensity changes. For simple phonemes, the time-structure can be divided into three parts even if they can be pronounced for arbitrary long time (of course within breath time). These parts are the begining of the pronouncation on-set, the clear steady phase of the phoneme, and the final part decay. Every begining lasts until the phoneme's maximal amplitude is not reached. Some phonemes have combined structures even if they are simple phonemes. E. g. the z and zs have special time-structures because the noise and the voiced sound arise at different time instants.

The connection of phonemes has more time-structure features than simple phonemes, because at the connection points, the complete structure changes in an attempt to provide the smooth transition from the previous to the following one.

So far we dealt with such attributes of the time-structure which are independent of the speaker and are evaluated subconsciously. This is called segmental time-structure.

The time-structure of the continuous speech is complex. A new part appears here, the break. Other partitions appear in the time-structure when we stress, change the rythm, emotionally prolonge the speech etc. These features, called supra segmental time-structure features, build up to the segmental structure. It is clear from the preceding that for a deaf child, the most difficult and critical phase of learning is to form the correct segmental structure because this structure is almost independent from our wishes.

The frequency- and intensity-structure are referred together as the spectrum of speech, comprising the relatively high energy frequency bands (the so called formant frequencies), their bandwhidths and the dim frequencies of the noise bands. So the spectrum gives information about all frequency components of the speech.

We are interested in the sound's pitch, intensity and timbre, which together form the above described spectrum. We also need to follow the changes of utterance. A suitable instrument can be a band-pass filter followed by a rectifier and a low-pass filter (Flanagan 1965.). Sweeping the centre of the band-pass filter through the frequency range of the signal causes the spectrum of the signal to appear at the output of the low-pass filter. This method of signal analysis is employed in the sound spectograph, or sonograph, the instrument used to obtain the sonograms [12]. This "voice print" is the sonogram, the originally three dimensional picture of speech in two dimensions. The horizontal and the vertical axis refer to the time and frequency, respectively, while intensity is marked with the darker light shadowing on the paper.

2.2. Instruments for teaching deaf children

In Hungary, about 2–3 in a thousand of school children are deaf and about 5–6% are speech impaired. Recently two Siemens Speech Monitors (SI 80-1) are used for the purpose to teach them (apart from the earphones and microphones used in speech laboratories). The Speech Monitor displays the pitch and the speech dynamics. In the case of training, the child tries to pronounce correctly a desired word. His sonogram is displayed on the lower half of the screen, while the reference pronounciation (the desired word's sonogram) can be seen on the upper half. The teacher or the child (depending on his age) can decide the further training by comparing the two sonograms. One problem is that this instrument does not give complete information about the speech even if it is in real time. Another disadvantage is the high price, 20.000 DM in 1989. There are other instruments for speech training but they indicate only some features of speech: Nasal-Indikator, f_0-Indikator, Sigmatrainer (to pronounce s, sch, ch). Other high quality instruments are the computer-controlled Speech Trainer by Siemens which uses an isolated word recognizer system to train the child to pronounce the desired word correctly, and Total Communication Workstation by English researchers which works similarly plus imitates the motion of the lips and synthesizes the word too [2] [4] [5]. In Section 2 we tried to emphasize the difficulties of deafness and show an instrument which examines the speech in its full. The method to make the speech visible hopefully can be applied to teach deaf children to speak. In the next Section, a possible realization on PC is shown.

3. A DIGITAL SPECTRUM ANALIZER

By digital filtering, a new output signal flow is derived from the incoming signal flow. There is a strictly defined polynomial connection between the two and it would be practical to find a simple relation between the input and output signal flow which can be handled easily. The steps of design and realization would become simple if we found building elements for which the transfer function was unambiguously expressed by a mathematical formula. The Wave Digital Filters meet these requirements, and have following advantages: automatically direct filters, economical solution of the hardware aspect, stop and pass band characteristics mutually complement each other, the signal rate can be economically reduced by a factor of two.

3.1. The wave digital filter (WDF)

Microwave systems are characterized by incident and reflected waves, and not only the transmission but the reflection can also be a transfer function. The "automatically direct filter" property means that every signal is tranfered which is not reflected totally [6] [7] [8] [9]. Consider a WDF two-port N (Fig.1.)

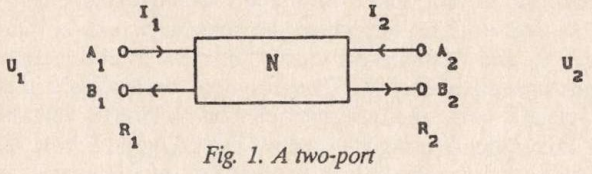


Fig. 1. A two-port

Where R_i are the port-resistances, $A_i = U_i + R_i I_i$ and $B_i = U_i - R_i I_i$ are the incident and reflected waves, respectively. Let the scattering matrix, S, have the following form:

$$\mathbf{S} = \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix}$$

If N is symmetric:

$$S_{11} = S_{22} \quad \text{and} \quad S_{12} = S_{21} \tag{1}$$

then

$$S_{11} = \frac{1}{2}(S_1 + S_2) \quad \text{and} \quad S_{21} = \frac{1}{2}(S_1 - S_2) \tag{2}$$

In the case of real frequences and passivity, the Feldtkeller-equation holds:

$$|S_{11}|^2 + |S_{21}|^2 = 1 = \det\mathbf{S} \tag{3}$$

It can be seen that if S_{11} denotes a high pass function then S_{21} becomes a low pass function, and vice versa. These kinds of reference analogue filters form the so-called lattice Wave Digital Filters. We obtain the transfer function of the digital filters by substituting $p = \frac{z-1}{z+1}$ into (2):

$$S_{21}(z^{-1}) = \frac{1}{2}[S_1(p) - S_2(p)]_{p=\frac{z-1}{z+1}} = \frac{1}{2}\left[S_1'(z^{-1}) - S_2'(z^{-1}\right] \tag{4.a}$$

$$S_{11}(z^{-1}) = \frac{1}{2}[S_1(p) + S_2(p)]_{p=\frac{z-1}{z+1}} = \frac{1}{2}\left[S_1'(z^{-1}) + S_2'(z^{-1})\right] \tag{4.b}$$

The most important advantage of this filter structure is the good scaling because the gain factor of all-pass filters is unity at all frequencies. Realization of the filters can be achieved by cascading at most second degree all-pass filters.

3.2. Bireciprocal filters

Consider a pair of filters having folowing transfer functions:

$$S'_{21}(z^{-1}) = A(z^{-2}) - z^{-1}B(z^{-2}) \qquad (5.a)$$

$$S'_{11}(z^{-1}) = A(z^{-2}) + z^{-1}B(z^{-2}) \qquad (5.b)$$

These transfer functions are related to frequencies f and $Fs/2 - f$ (Fs is the sampling frequency). Let

$$G(f) = S'_{21}(z^{-1})_{z=e^{j2\pi f/Fs}}$$

and

$$\hat{G}(f) = S'_{11}(z^{-1})_{z=e^{j2\pi f/Fs}}$$

At frequency $Fs/2 - f$,

$$G(\frac{Fs}{2} - f) =$$
$$= A(z^{-2}) - z^{-1}B(z^{-2})|_{z=e^{j2\pi(Fs/2-f)/Fs}} =$$
$$= \hat{G}(-f) = \hat{G}^*(f),$$

where * means the complex conjugate value. If $G(f)$ denotes a high-pass filter then $\hat{G}(f)$ becomes low-pass filter and vica versa, as we have seen in (3).

Moreover, if A and B are all-pass filters, than it can be seen from (3) that

$$|G(f)|^2 + |\hat{G}(f)|^2 = 1, \quad \text{if} \quad f \in (-\frac{Fs}{2}, \frac{Fs}{2})$$

and

$$|G(f)|^2 + |\hat{G}(\frac{Fs}{2} - f)|^2 = 1, \quad \text{if} \quad f \in (-\frac{Fs}{2}, \frac{Fs}{2}).$$

Therefore we call the

$$S_{21}(z^{-1}) = \frac{1}{2} \cdot \left[S'_1(z^{-2}) - z^{-1}S'_2(z^{-2})\right]$$

$$S_{11}(z^{-1}) = \frac{1}{2} \cdot \left[S'_1(z^{-2}) + z^{-1}S'_2(z^{-2})\right]$$

pair of filters a bireciprocal filter pair due to their dual symmetrical properties.

The bireciprocal property has a comparatively strict specification for the zeroes and poles of the transfer function but at the same time, the realization of the filters is easy, thus making their application attractive.

The output signals can be decimated with relatively small error if the transition bands of the filters are narrow. This is useful if we need a filter bank comprising filters having the same passband. This kind of bireciprocal filter structure is shown in Fig. 2, using all-pass functions for S_1 and S_2. Their sum and difference yield the high pass function S_{11} and the low-pass function S_{21}, respectively.

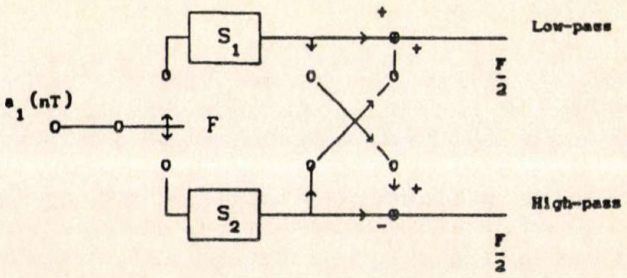


Fig. 2. A bireciprocal structure

The high-passed and low-passed frequency products arise in the lower and upper branches, respectively. These filtered products must be logarithmized to represent the energy of the signal in dB in all appropriate frequency bands. If further resolution is required, a second stage must be coupled into the lower and upper branch.

We can see the advantages of the wave digital filters from this arrangement: it yields the filtered products directly, the transfer characterictics are symmetrical to $F_{sample}/4$, the sampling rate in the inner branches is the half of the original one, thus decimation can be achieved and a number of elements can be economized in the realization.

3.3. Bireciprocal filter bank with TMS 32010

The 0–4 kHz speech band can be split into 32 consecutive bands with the same 125 Hz bandwidths. In the filter bank the sampling rate is halved in every stage while the number of the all-pass filter modules is doubled. For obtaining equal effective bandwidths, different specifications must be used in every stage. For all 32 band-filters, about 35 dB attenuation can be achieved by applying 9th–9th–7th–5th–5th degree Cauer filters in the consecutive five stages.

The filtering process is carried out by the TMS 32010 microprocessor which is able to perform fast multiplying, addition and shifting simultaneously. This microprocessor uses different 16 bit buses for the purpose of moving data and fetching commands. The TMS has the following features: 200 nsec cyclic time, 144 words inner RAM, 32 bits ALU and accumulator, programmable shift register. The outer data RAM is divided into two pieces which are used either by the TMS or by the PC at the same time. Each RAM slice has 2 kword capacity.

We applied the folowing algorithm. The sampled signals arrive from the Analog/Digital converter to the TMS's inner RAM at the sampling frequency because the analog/digital converter is forced to take in a sample from the microphone every 125 μsec. The TMS executes the filtering procedure on 32 consecutive samples throught the five stages of the filter bank. We save the actual states of all the delay units in the filter bank instead of saving every time the input and output of each filter block. In this way, the resulting filtered outputs can be stored on the same place of the original 32 samples and the processing speed increases, thus it is possible to display the results by the PC in every 12 msec.

4.APPLICATION

Before we introduce a real-time digital spectrograph, it should be explained why *wave digital filters* are applied instead of a *Fast Fourier Transform algorithm.* First, when we decided to make a real-time spectrum analyzer, we wanted to use a method which is fast enough and does not require much memory. Secondly, we were interested to try a new algorithm hoping it will be suitable for our purposes and will give a novel approach in filtering.

This 32-band spectrum analyzer provides a robust resolution in the 0—4 kHz frequency band, thus in a possible application, it is would be sufficient to examine this rather narrow speech band, and the displayed results (32 parameters) would reveal for the user the main characteristic of speech. We chose the display method of the spectrograph, and we tried to expand the field of the application. Thus we found the speech training idea: children try to pronounce a word correctly and the pronounced sonogram is compared with a reference one. Similar sonograms would hopefully mean understandable speech.

4.1. Graphics

How does the graphical display work and how can it be accelerated?

The PC inputs a 32 element integer segment from the TMS every 12 msec. These incoming data are proportional to the absolute value of the sampled speech signals so we have to convert these to a dB scale by taking 20 $\log_{10}$. These results are classified according to a 5 section scale where the indices of the scale are the row numbers of the predefined images. These images represent higher intensity by darker display (more highlighted pixels). To obtain a fast display, we defined a logarithmic table at the beginning of the program and also produced the five types of the displayable images. In this way, the incoming magnitude-type data address the images which are then displayed on the appropriate position of the sonogram (at given time and frequency band). The display occurs in two part simultaneously: on the lower and bigger part of the screen the sonogram can be seen (time—frequency—intensity) while on the upper and smaller part, the total energy curve (time—intensity) is displayed continously. The display is illustrated in Figure 3.

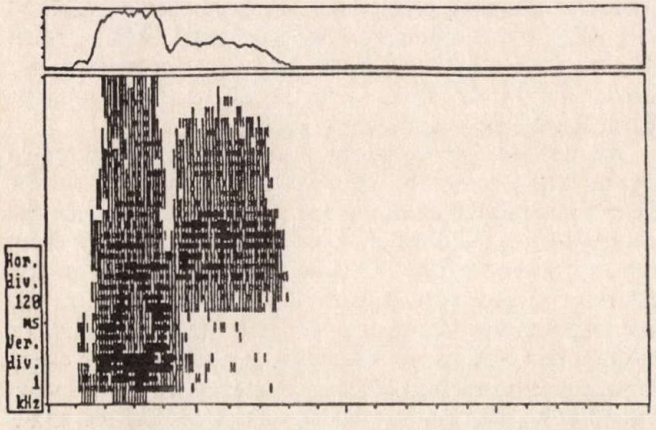


Fig. 3. The sonogram and intensity curve of the word "vajas"

4.2. Software for speech training

The program is guided with pop-up menus and hot keys to provide easy operation [10]. The program has an editor and a training section. First we have to store reference words in a pattern dictionary for the later training. The teacher or the phoniater pronounces a word which is displayed on the screen. If she or he finds it suitable for training, the sonogram can be saved by the hard disk. Before saving it, we can write an additional text for the sonogram. This subtitle is displayed on the lowest row of the sonogram, and subsequently, the sonogram with the text can be saved, together with an arbiratry designation.

The software offers some features to make the application easy for the user. Thus we can put vertical and horizontal grids on the sonogram, or we can see the spectrum at an arbiratry time point projecting from the two dimensional sonogram the height (intensity) information. This is called amplitude slice (see Fig. 4.) and can be moved by the cursor over the total 2.4 sec interval of speech.

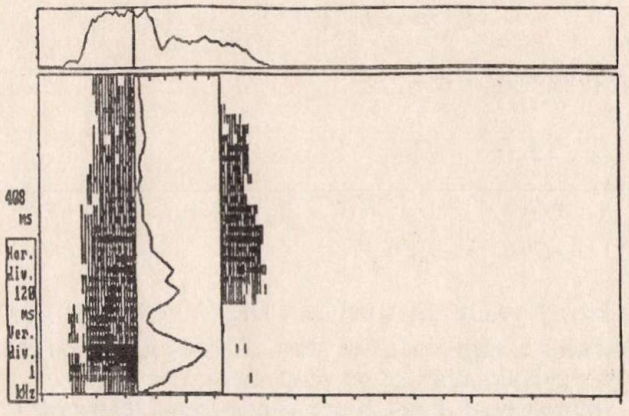


Fig. 4. Amplitude slice at voice 'a' in the word "vajas"

The TMS provides results continuously. To recognize useful information, namely speech for displaying, the software uses automatic beginning detection which means that displaying occurs only when the total incoming energy exceeds a predefined limit over some consecutive 12 msec.

In the training section, the screen is divided into two parts. The teacher loads a previously saved reference word's sonogram to the upper half. The child tries to

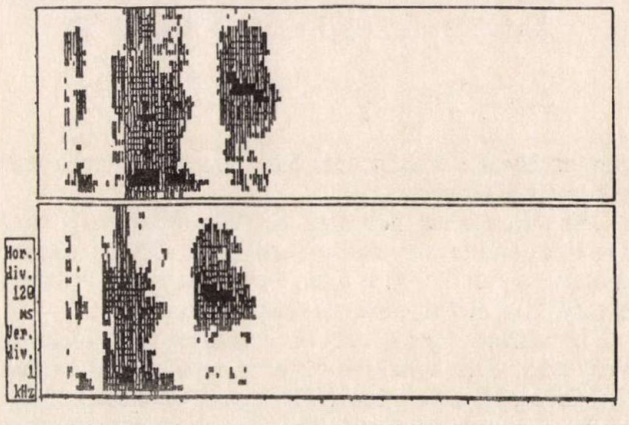


Fig. 5. The training with word "kikerics"

pronounce the original word correctly, and his utterance is displayed on the lower part of the screen. The two sonograms can be compared (as in Fig.5.) depending on the age and the ability of the child with or without the help of the teacher. The similarity of the two pictures decides the further training. It should be mentioned that two sonograms will never be the same because no one can produce exactly the same speech waves two times. Thus the two sonograms can show similarity as it is clear in Figure 5 where one of the authors produced the two versions of the word "kikerics".

4.3. Conclusions

Finally, we should emphasize that the result of this work gives only quantitative information about the speech. The most important advantage is the real-time processing and displaying which can show the speech in its whole. The children are able to follow instantly what happens if they change the rythm, the pitch or the intensity of their speech. In the future, we would like to add more features of an "instrument" to the equipment, and develop some additional features like displaying the pitch and the format trajectory. The pitch display can be useful to teach the children to pronounce correctly the suprasegmental structure while the format trajectory will show much more information about the basic speech product.

5. SUMMARY

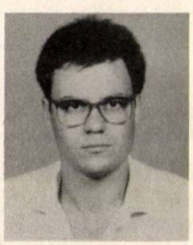
A real-time application of the wave digital filter bank was introduced in the paper. We showed the basic properties of this filter type and proved an economical decimation technique. As a hopefully useful application, additional software for teaching deaf chlidren to speak was also presented. Some remarks about further possible development were finally discussed.

ACKNOWLEDGEMENT

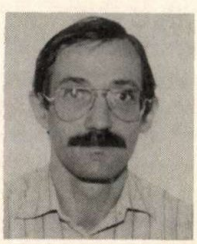
The authors are deeply grateful to Dr. G. Gordos, head of the Department of Telecommunications and Telematics, for his constant help and encouragement.

REFERENCES

[1] Olaszy, G., "Electronic speach generation" (in Hungarian), Műszaki Könyvkiadó, Budapest, 1989.
[2] Parsons, W., "Voice and Speech Processing", McGraw-Hill Book Company, 1986.
[3] Siemens, "Speech Monitor SI 80–I" and "Computer-Controlled Speech Trainer" prospectus.
[4] Kingham–Harris–Tolmie, "The Integration of Speech Technology with Graphics as an Aid for the Disabled", *Europian Conference on Speech Technologie*, Edinburgh, 1987.
[5] Maassen, Arends, Povel, "Artificial Corrections to Deaf Speech and Development of Visual Speech Training Aids", *Europian Conference on Speech Technologie*, Edinburgh, 1987.
[6] Géher, K., "Linear Networks" (in Hungarian), Műszaki Könyvkiadó, Budapest, 1972., pp. 142.
[7] Wegener, W., "Wave Digital Directional Filters with Reduced Number of Multipliers and Adders", AEÜ, Band 33, 1979., pp. 239–243.
[8] Fettweiss, Nossek, Meerkötter, "Reconstruction of Signals after Filtering and Sampling Rate Reduction", *IEEE Transactions on Acoustics, Speech, Signal Processing*, vol. ASSP–33. Aug. 1985., pp. 893–901.
[9] Gazsi, L., "Explicit Formulas for Lattice WDFs", *IEEE Transactions on Circuits and Systems*, vol. cas-32., Jan. 1985., pp. 68–88.
[10] Boda, P. P., "Spectrum Analyzer Development with WD Filter Bank" (in Hungarian), *Thesis work*, TUB, 1991.
[11] Gordos, G., Takács, Gy., "Digital Speech Processing" (in Hungarian), Műszaki Könyvkiadó, Budapest, 1983., pp. 15–63.
[12] Ainsworth, W. A., "Speech Recognition by Machine", Peter Peregrinus Ltd., London, 1988., pp. 56.



Péter Pál Boda graduated in electrical engineering at the Technical University of Budapest in 1991. He is currently pursuing the Ph. D. degree in electrical engineering at the Department of Telecommunications and Telematics, TUB. His research interests include speech processing and neural networks. He is a student member of the IEEE Signal Processing Society.



László Osváth graduated in electrical engineering in 1975 at the Technical University of Budapest. After a period of postgraduate studies, he joined the TUB in 1976 where he is an assistant professor in the Department of Telecommunications and Telematics. His principal interest is digital signal processing, mainly it's application in data transmission and speech processing.

MULTILINGUAL TEXT-TO-SPEECH CONVERTER*

G. OLASZY

PHONETICS LABORATORY,
LINGUISTIC INSTITUTE OF THE HUNGARIAN ACADEMY OF SCIENCES,
BUDAPEST, HUNGARY

G. NÉMETH

DEPARTMENT OF TELECOMMUNICATIONS AND TELEMATICS
TECHNICAL UNIVERSITY OF BUDAPEST
H-1111 BUDAPEST, STOCZEK U. 2.

A general, multilingual text-to-speech (TTS) technology has been developed in Hungary for Europe. The commercialised version of this research is the MULTIVOX system. It automatically converts Hungarian, Finnish, Dutch, Italian Spanish and Esperanto text into standard speech (with rhitm and intonation) in real time. A Standard Arabic version of this system is also available. Two main operating modes can be used: (i) automatic reading of text files (in ASCII form), and (ii) programming of unrestricted speech events (messages, information, task evaluation by voice, warnings, etc.) into any program. The system consists of a speech synthesizer box of small size, a resident program for the TTS synthesis and a computer. The resident program can be reached from high level languages (C, Pascal, Turbo-Pascal) and assembler. It runs under DOS, and Windows. The phonetic aspects of the system, the main building blocks, their operational principle, the main steps of adaptation to a new language are described.

1. INTRODUCTION

Phonetics is a traditional science in Hungary. The investigation of acoustic and linguistic features of speech sounds and speech began already the begenning of this century and is continued today at PhL HAS, the leading phonetics research center in Hungary. The first research, to create a Hungarian text-to-speech (TTS) synthesis system was performed in the late 70's in the PhL HAS. An analysis/synthesis system was designed and implemented on a DEC PDP-11 computer and an OVE III freely programmable format synthesizer. The first demonstration of the Hungarian speaking PDP-11 computer took place at the First Conference of PDP Users, Budapest, 1981.

On the basis of this research at PhL HAS, two research groups began the work in the early eighties for further specialised developments in speech processing.:

- Speech Lab. at TUB started a cooperation with PhL HAS for further linguistic and technical development for speech synthesis techniques, concentrated on applications for PCs (Commodore 64, later on IBM PC machines) and for stand alone synthesizers.
- Central Research Institute for Physics (CRIP) initiated another cooperation with PhL HAS which resulted in the transfer of a PDP controlled MEA8000 format synthesizer chip technique (developed at PhL HAS) to them. This transfer and later personal help and consultation formed the basis of the development of specialized speaking computers for the blind (BraiLab) in CRIP. These developments were continued later without the paricipation of PhL HAS.

The latest result of the intensive 10 year research of developement at DTT TUB and PhL HAS is the MULTIVOX multi-lingual text-to speech philosophy and technique. The commercialized realization of this appeared in several HW and SW solutions like small size, easy-to-use, programmable and inexpensive multilingual TTS, and also limited vocabulary synthesizers for general use in industry, education, research and medical applications. A new portable speaking aid for the speech impared (PORTTalker) has also been developed. During this long research period, the general philosophy of the original solution has been redesigned. A new development system, MULTIDEV, has been created as well, to ensure a flexibile and efficient environment for the developement and the adaptation of new languages. The PC based version of the latest system (MULTIVOX) consists of a Centronics interfaced standalone box and a SW. The SW was developed as a terminate-and-stay-resident program running under DOS or as a dinamically linkable and loadable (DLL) module under Windows. The speciality of the MULTIVOX system is its multilingual feature. It can be used for conversion of any written text – with no limitation in content and time – in 8 languages (Hungarian, Finnish, German, Dutch, Italian, Esperanto, Spanish, Standard Arabic). Further languages are under development. The system is based on the original idea of acoustic building units (ABU) which greatly support multilingual synthesis.

2. THE BUILDING ELEMENTS OF THE MULTIVOX SYSTEM

The general blocks of different TTS converters are mainly the same. Every system must have a grapheme-phoneme converter, an acoustic representation of the given sounds, a prosody module etc.

From the point of view of the phonetic and technical implementation, the only thing that make the systems different are the inner solutions, the language dependent modules, the hardware, and the aim of making a final commercial product or not. In this respect the MULTIVOX system (Fig. 1) had several new solutions, especialy from the point of view of multilingual operation.

MULTIVOX has both language dependent blocks (marked with dotted line), the content of which are redesigned when adapting the system to a new language, and language independent parts, which are the same for every language. The bidirectional arrows in Figure 1 refer to the interdependence of certain blocks, which has to be considered during system development. This unique system design philosophy makes it possible to adapt the system to different languages. The adaptation process takes 3–6 month for a new language.

3. MULTILINGUAL TEXT PREPROCESSING

In a multilingual approach, a text preprocessing module is needed on the grapheme level to handle the different letters of different languages. A unified grapheme repre-

* This is a revised and abbreviated version of a long paper to appear in G. Bailly, C. Benoit (ads.): Talking Machines: Theories, Models and Applications, Elsevier

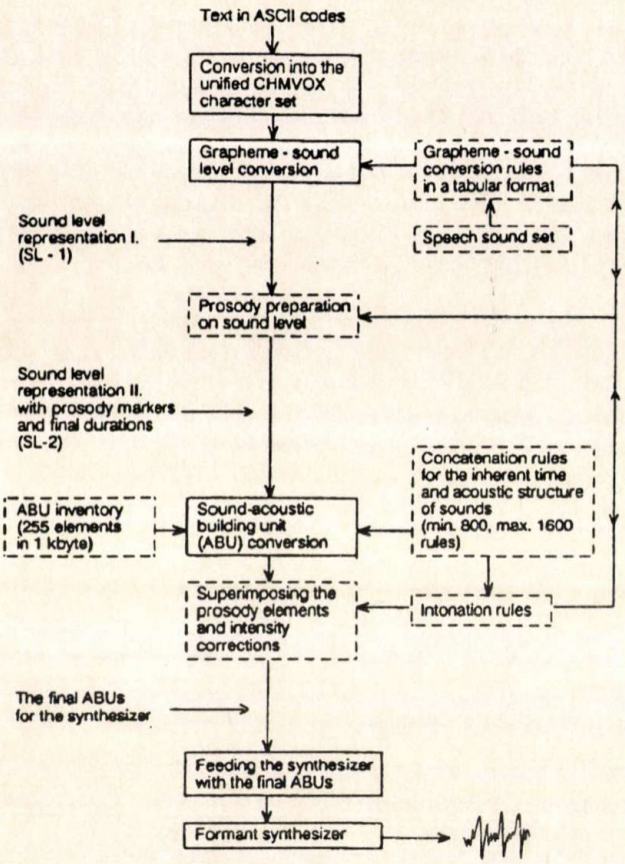


Fig. 1. The main building blocks of the MULTIVOX system

sentation, the CHMVOX (CHaracters in MultiVOX) character set, was designed for the general description of letters with diacritics in different languages. A filter in the preprocessing stage of text ensures the correct conversion of these letters into the unified CHMVOX character set. In the CHMVOX representation, every letter with a diacritic is represented by two characters: the letter itself and the representative of the diacritic (for example á=a', ó=o'). Three filters are available for the user, corresponding to the IBM ASCII, the ANSI and the ROMAN8 character sets. The special letters with diacritics, their ASCII codes, and their CHMVOX representation can be seen in Table 1.

The different character sets can be chosen either in a configuration file or by software options. This CHMVOX form is applied for the internal character description in the MULTIVOX system, so it can always be used for text input (for example it is acceptable to type either Hölle or Ho:lle for 'hell' in German). The text of any language will always be transformed into CHMVOX characters, regardless of whether it was given in IBM-ASII, ROMAN8, or ANSI characters. Upper case letters, after having been marked for future purposes, will be converted into lower case ones.

Numbers are treated separately and converted into their written form. Separate rules serve to solve this conversion for every language. The system is designed to handle numbers up to 1 billion.

Most common abbreviations and acronyms are converted into their spoken form. They are stored in a vocabulary which can be extended by the user, so field specific abbreviations and acronyms can be built into the system.

Letter	IBM-ASCII		ROMAN8		ANSI		CHMVOX		Language*					
	upper	lower	upper	lower	upper	lower	upper	lower	H	F	G	D	I	E S
ä	142	132	216	204	196	228	A:	a:		x	x			
ö	153	148	218	206	214	246	O:	o:	x	x	x			
ő		147	223	194	212	244	O"	o"	x					
ü	154	129	219	207	220	252	U:	u:	x		x			
ű		150	174	195	219	251	U"	u"	x					
ß		245	222	223			B				x			
á		160	224	196	193	225	A'	a'	x					
é		162	231	198	211	243	O'	o'	x					
ú		163	237	199	218	250	U'	u'	x					
í		161	229	213	205	237	I'	i'	x					
é	144	130	220	197	201	233	E'	e'	x					
ù		151						u`				x		
à		133						a`				x		
ò		149						o`				x		
ì		141						i`				x		
è		138						e`				x		
š							S^	s^						x
č							C^	c^						x
ĵ							J^	j^						x
ĝ							G^	g^						x
ĥ							H^	h^						x
ñ	165	164					N^	n~						x

* H=Hungarian; F=Finnish; G=German; D=Dutch; I=Italian; E=Esperanto; S=Spanish
These abbreviations are used throughout this article.

Table 1. Handling of special ortographic characters

4. SPEECH SOUND SET FOR MULTILINGUAL SYNTHESIS

A general collection of speech sounds (Table 2) is the basis of the text-sound-code transformation (in practicle this is called grapheme-phoneme conversion).

Sound category	Speech sounds
Vowels	a, o, o:, ɔ, u, y, i, ɐ, ü, ü:, æ, ɛ, e
Stops	b, p, d, t, g, k, j, c
Nasals	m, n, ɲ, ŋ
Fricatives	v, f, h, ɦ, j, ç, x, ʁ, z, ʃ, ʒ
Affricates	ʦ, ʧ, ʤ, ʤ
Lateral	l
Rolled	r, ʁ
Diphthongs	ɔ͡i, a͡u, o͡u, e͡i, ɐ͡i, o͡y, a͡y
Specials	ŋ̍, kv̍, gv̍, ə1, ə2, ɐ̯

Table 2. Speech sounds of the MULTIVOX system

The total number of speech sounds used for generating the languages mentioned above is 56. The sounds in this table represent only phonological level theoretical units (phonemes), the phonetic level variants of the same phoneme are implemented separately for every language. The maximum number of speech sounds allowed for one language in the MULTIVOX system is 40. This means that a maximum of 40 sounds from Table 2 can be used for the synthesis of the language. This is an empirical value that resulted from the experience of synthesizing languages in the last decade. In Table 3 speech sounds and their codes (distibution) are shown according to the seven fully implemented languages. As it can be seen, speech sounds are numbered for internal processing. Inside the system, these code-numbers refer to the sounds.

When adapting the system to a new language the first step is to determine which of the already existing source languages will be modified to produce the new target language. It can be seen from the organization of the speech sound set table, that the sound representation

of Hungarian (the first language developed) basically influenced the distribution of sounds of other languages, The closer the target language was to an already working one the easier was the developement. For example, Italian was developed from Esperanto, Spanish from Italian and Dutch from German.

The general rule for the placement of a target language sound in this table is as follows: those sounds from the target language that were also represented in the source language will be assigned the same code number, and the rest of the target language sounds will be assigned to the remaining places in the table. This organization is only a phonemic level one, on the phonetic level, the acoustic data of the speech sounds having the same code number in the table may be different. This type of organization for assigning the target sounds is one of the most important factors in the quick and effective adaptation of a new language. This ensures that the majority of the sound combination rules of the source language can be also used for the target language without much correction.

Year Lang.	Speech sounds and their code numbers
	1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40
80-89 H.	- a ɔ o u y i e: ø e b p d t g k ʒ c m n ɲ j h v f z s ʃ ʒ ʃ ʧ l r
84-90 G.	- a ɔj ɔ u y i e ø e b p d t g k ɛ ə l m n ŋ j h v f z s ʃ ʒ ʃ ʧ l R ər əu ei ç x o ə2
86-90 F.	- - a o u u i e b p d t g k ʤ ɲ m n ɲ j h v f z s ʃ ʒ ʃ ʧ l r
88-89 F.	- - a o u y i e ø æ b p d t g k h m n j h v f z s ʃ ʃ l r
89-91 I.	- a ɔ o u gv i kv ɛ e b p d t g k ʤ ʃ j m n ɲ j ʤ v f z s ʃ u ʃ ʃ ʧ l r
90-91 D.	- a: a o u y i ei ø ou b p d t g k e ə m n ŋ j h v f z s ʃ ʒ ʃ ʧ l r v au ei x ø y ʌ/ ɐ
91-92 S.	- a - o u gv i kv ɛ - b p d t g k ʃ j m n ɲ j ɲ v f z s - u ʃ ʃ ʧ l r

Table 3. Speech sounds distribution in the MULTIVOX system

5. CONVERSION OF GRAPHEMES INTO SPEECH SOUND CODES

The written form of a language differs from the spoken one. Therefore every text must be converted into sound level representation. This conversion is language dependent. A special, tabular description, the MULTIVOX table (MT), is used for this purpose.

The main task of the MT is to transform the letters of the text into sound level information i.e. into the sound codes according to Table 3. Two additional subtasks are fulfilled here: (i) setting timing inside words (sound durations, short/long oppositions, glottal stop, vowel reduction, etc.), and (ii) to prepare prosody and intonation processing by placing special markers in the rules (e.g. the sequence is unaccented, it is joined to the following word, it is a question word, etc.). The MT does not handle numbers.

The syntax of the rules in the MT is simple and language independent, only the content of the rules depends on the language. The form of the rule is as follows: the left side contains the letter sequence (L SEQ), while the right side contains the sound codes (SND) of the spoken form of the sequence.

L SEQ + pointer setting = SND + markers for prosody (1)

The rules are alphabetized according to the first letter of the sequence. Each letter of the alphabet represents a separate rule block in the table. The organisation of one such block is similar to a saw-tooth triangle contour, which is the widest at the top of the block where the longest rule is placed (having the highest number of the characters in the rule). The peak of the triangle is at the bottom i.e. the last rule consists of only one letter (the initial letter of the unit itself). An example of the organisation of blocks beginning with letters c and d (from the German MT) are shown in Table 4.

chor = 16 4 4 4 4 33 0 <—	demonst = 13 17 17 19 4 4 20 30 14 0
ckel = 16 16 18 32 0	deutsch = 13 3 31 0
cho = 37 4 4 0	dialekt = 13 7 2 2 32 17 17 16 14 0
cyr = 28 6 6 33 0	dachsp = 13 2 2 38 30 12 0
ch = 38 0	dachst = 13 2 2 38 30 14 0
ck = 16 0	dament = 13 2 2 19 18 20 14 0
c = 16 0	dete\ = 13 17 17 14 17 17 0
	dabe\ = 13 2 2 11 0
	dachs = 13 2 2 16 27 0
	dame\ = 13 2 2 2 2 19 0 <—
	darf\ = 13 2 2 33 0
	datt\ = 13 2 2 14 0
	dware = 13 24 10 10 10 33 0 <—
	dar\ = 13 2 2 2 2 0
	dat\ = 13 2 2 0
	da: = 13 10 10 0
	dau = 13 35 0
	div = 13 7 7 24 0
	da = 13 2 2 0
	d = 13 0

Table 4. The c and the d rule blocks of the German MT

5.1.1. Setting the basic sound durations in the MULTIVOX Table

The duration of a sound is characterized in the MT by using a theorethical unit. By prescribing (concatenating) more units, the duration of the sound can be set increasingly longer and vice versa. The physical duration of one unit is shorter (20—40 ms) than a phoneme duration and is set exactly at SL-2 level. The symbol of the unit is the sound code of the letter combination. The MT rule designer has the freedom to choose the number of concateaned units. This type of organization allows the designer to prescribe shorter and longer variants of a phonologically short or long vowel already in the grapheme-to-sound code conversion stage of the process. In the MT, a normal short vowel is expressed by two sound units. Other durational variants can be expressed by using more or fewer units, as in this example (V being one unit):

Shorter variant	normal duration	longer variants	long versions
V	VV	VVV	VVVVV

The timing units for consonants were designed separately for each consonant. Two types of solutions are used to describe the sound durations: setting the sound durations according to linguistic rules, and handling the exceptions to these rules. In the MT the designer can prescribe the basic duration of sounds in terms of short and long. There are two possibilities in indicating the duration of a sound: (a) It is exactly indicated in the written form of the language (like in Hungarian, Finnish, or in some cases in German, Italian and Spanish) or general rules can be applied (like the penultimate syllable lengthening rule in Esperanto or Italian in open syllable) or (b) specific linguistic rules have to be used to prescribe whether a sound must be short or long (German, Italian, Spanish etc).

The rules of case (a) can be expressed exactly in the MT. For example, in Hungarian the long versions of sounds are written either by doubling the letter of the sound (for consonants) or by using a diacritic sign (for vowels). In German some special letter combinations indicate certain lengthenings (like $ie = i:$ *or* $ah = a:$ etc) as well. As for case (b) rules, morphological analysis is a proper way to determine whether a vowel must be long or not [26]. No morphological analyzer is used in MULTIVOX, because of memory and speed limitations. Special letter sequence rules were developed and are described to realise the correct lengthning.

5.1.2. Exceptions in lengthening

In some cases, the letters indicating a long sound are not pronounced as a long one (e.g. $i' = [i]$, $u' = [u]$) and vice versa (e.g. $j = [j:]$; $t = [t:]$). Examples for Hungarian and German are given in Table 5. In these cases, the designer puts the exceptions directly into the MT.

Written form	Spelling pronounciation	Actual pronounciation	Language	Gloss
szinház	[si:nha:z]	[sinha:z]	H	'theatre'
újság	[u:ja:g]	[uja:g]	H	'newspaper'
legújabb	[legu:jab:]	[leguj:ab:]	H	'the newest'
egy	[ej]	[ej:]	H	'one'
Folie	[fo:li:]	[fo:lie]	G	'foil'
gehören	[ge:ø:ren]	[gehø:ren]	G	'to belong to'

Table 5. Differences between written and spoken forms

5.2. Second level processing of sound codes

In the step (SL-2) timing modifications are made according to the rules of the prosody preparation module. These modifications are language dependent and mainly concern:

- the preprocessing of vowel durations for word accentuation (in languages where the accent co-occurs with lengthening),
- The realization of vowel reduction phenomena (if it exists in the given language),
- the creation of a glottal stop if necessary,
- the performance of durational changes in one-word questions, and
- the placing of markers in the sound code series for future timing intonation processing.

Word	First step in SL-1	First step in SL-2	Remark	Language
mama	maamaa	maaaaaama	open syllable	I
quattro	quaattroo	quaaaattro	closed syllable	I
fara'	faaraa	faaraaa	final accent	I
(m)nella	neellaa	neella	no accent	I
esperanto	eespeeraantoo	eespeeraaantoo	closed syllable	E
facila	faaciilaa	faaciiiiilaa	open syllable	E
(m)preter	preeteer	preeteer	no accent	E

Table 6. Word accent preparation at sound levels SL-1 and SL-2 ({m}=marker)

5.2.1 Word accent

The preprocessing of word accent concerns Italian, Spanish and Esperanto, where the lengthening of the accented vowel in open syllables co-occurs with accent. In these languages penultimate accent is a general rule. In Italian and Spanish there are some exceptions to this rule, so that accent can be on other syllables of the word as well. To handle these cases separate rules were designed.

5.2.2. Exceptions in accentuation

An important function of the MULTIVOX TTS system is that it handles also those cases in which the general accentuation rules must not be used. These exceptions (for Italian and Esperanto) are handled in the MT by marking them with the "no lengthening" marker. Such words include in Esperanto the so called "table-words" like: kio, kiu, kie, kiam, kial, kioma etc., the possessive pronouns, and the prepositions; in Italian the two syllable articles like nelle, della, etc. The pronounciation of these words is performed only with normal short vowels.

5.2.3. Vowel reduction

To imitate vowel reduction, only one sound symbol is left for the vowel in question. The word final reduction algorithm for Italian runs as follows:

$$if\ word\ final\ (VV)\ then\ reduce\ to\ (V). \quad (2)$$

The result of reduction (and of lenghtening rules discussed above) are shown in Table 7. As it can be seen, the extent of lengthening is different in the two languages. Our experience is that lenghtening has to be greater in Italian than in Esperanto. This is the case for Spanish, too.

Units expressing the duration of a vowel					Language
Neutral vowel	Accented vowel			Reduced vowel	
	open syllable	closed syllable	final position		
VV	VVVVVV	VVVV	VVV	V	I
VV	VVVVV	VVV	-	-	E

Table 7. Timing units for sound duration design

5.2.4. The glottal stop

A special phenomenon, the glottal stop, (marked with the diacritic sign' in the examples) is also handled by the module for prosody preparation. In German the initial vowels are preceded by a glottal stop. Even the addition of an unaccented prefix does not change this rule (antworten 'antworten, beantworten be'antworten). The glottal stop algorithm works only for the latter case. This means that

the algorithm of the prosody correction module inserts a marker between the two vowels. Physically, a short (about 20ms) pause will be put between the two vowels, and this results in the glottal phenomenon in the final speech signal.

5.2.5. Special durational change in questions

Experiments gave a new result, that in one-syllable questions, the duration of the vowel is longer than in the same word if it is pronounced as a declarative sentence (in Hungarian Már? and Már. 'already', Most? and Most. 'now'; in German Gut? and Gut. 'good') [19]. So the duration of the vowel depends on the sentence type. If a correct effect is to be achieved in both the question and the declarative form, then the duration must be set according to the sentence type. In the MULTIVOX system the default case is the declarative form duration and this value is lengthened in case of one-syllable questions.

Marker	Meaning	Lang.	Marker	Meaning	Lang.
45	the word is unstressed	all	70	join preposition	E, G
46	the word is neutral	all	71	join pers. pronoun	E
47	open syllable accent	E, I	72	first syllable accented 3	H
48	closed syllable accent	E, I	73	first syllable accented 4	H
49	the word is a question word	all	80	speed the word up	all
50	first syllable accented 1	H, F	81	slow the word down	all
51	first syllable accented 2	H	84	accent in the penult. vowel	G
52	pitch higher in the word	all	88	join the word backward	all
55	verbal prefix, accented	G	96	join unaccented words	G, I
56	verbal prefix, unaccented	G	97	accent in the last vowel	G, I
58	accent in two-syllable word	G	100	do not touch the word	G
60	the word will be louder	all	101	no glottal stop	G
61	the word will be softer	all	253	semicolon (340 ms pause)	all
64	pitch up with 8 Hz	all	254	comma (160 ms pause)	all
65	pitch down with 8 Hz	all	255	colon (530 ms pause)	all

Table 8. Markers for prosody and for options

5.2.6. Markers for prosody routines

Special markers (number above 40) are used (Table 8, Table 9) in the MT to mark those sound sequence points where prosody (or any other) changes are required. At this level (SL-2), the sound code representation already contains all duration and marker data for pronounciation. The intonation markers will be processed in a further step in the intonation module.

Text in characters	Speech sound codes	Gloss
a.	1 2 2 1	'a'
2.	1 28 24 36 1	'two'
Ich komme.	1 7 7 38 1 16 4 4 19 18 1	'I come'
Es ist gut.	45 17 17 27 7 7 27 14 52 15 5 5 5 5 5 14 1	'This is good.'
Wann kommen sie ?	49 24 2 2 2 20 16 4 4 19 19 17 20 1 24 7 7 7 7 1	'When do they come ?'

Table 9. Examples of final results of the letter-to-sound conversion in German

6. TRANSFORMATION OF SOUND CODES INTO PHONETIC UNITS

The output of the grapheme-phoneme conversion is the spoken form representation of the written text. The transformation of the sound level information into phonetic level units is a multilevel task. Two main modules are used in this process: the inventory of phonetic level Acoustic Building Units (ABUs), the rule system for concatenating the ABUs to get the code level representation of speech signal to be synthesized.

6.1 The acoustic level representation

In the MULTIVOX system, the phonetic data are stored in ABUs organised in an inventory, consisting of a maximum of 255 ABUs. Every language has its own ABU inventory. One ABU is described by 40 bits which prescribe the control parameters of the synthesizer (formants, bandwidths, duration amplitude, pitch increment, etc.) and their internal values. When designing the first ABU inventory in 1980–81 with 370 ABUs to describe Hungarian [16], the phonetic data for ABUs were obtained from sonograms [17]. The ABU inventory for subsequently developed languages was mainly determined by simply correcting these data on basis of phonetic descriptions concerning the language in question. Only completely new sounds were designed using the sonogram technology. Empirical calculations show that, during the adaptation for a new language, only 30-40% of the ABU inventory has to be redesigned. In every case, the development of ABUs was supported by the MULTIDEV interactive development system and by native speakers.

An important feature of the MULTIVOX system is that each ABU represents a short (in general much shorter than a phoneme) acoustic event of the speech signal. When designing these ABUs, an interpolation procedure was also taken into consideration. This means that the data of two ABUs are interpolated in the final process (Fig. 2.). Another special feature of the ABUs is that many of them were designed for general purpose use. The fact that the majority of ABUs have a shorter frame duration than a speech sound leads to the result that one ABU will not strictly represent a given speed sound or sound combination, but it may be used in all places of the speech wave where the desired acoustic content matches or closely matches the acoustic content of the given ABU. Only the ABUs representing the steady states of vowels and some consonants are exceptions to this concept.

The above described special kind of design results in highly reduced set of ABUs (maximum 255 types) representing the whole acoustic structure of the given language. The final set of the ABUs is a result of a long interactive development process where a repeated listening + studying + correcting procedure (for sounds, sound combinations, words, sentences) was the basic method of getting the final data for synthesis. Using and concatenating the proper ABUs, every sound part, sound, or sound combination of the language can be realized.

6.2 The rule-system for concatenation of ABUs

The sound code level description of a text results in only a theorethical phonemic-level determination. If a real speech signal is desired, this phonemic-level information must be converted into phonetic-level information in which the physical data for generating a speech signal will be present to control the synthesizer. The transformation of sound codes (SL-2) into ABUs is performed by the concatenation rule (matrix) system. The input data of the rule system are the output data of the letter-to-sound code

converter module. Six to ten ABUs must be concatenated to describe and create the control codes for a speech sound or a sound combination. Every sound combination of the given language can be realized by using appropriate ABUs.

Row	Column	Rule	Rule concatenated ABUs	Remark
1	23	-H	225,226,0,0,0,0	225, 226: parts of [h]
23	4	HO	233,3,23,23,0,0	233, 3, 23: parts of [o]
4	4	OO	23,23,0,0,0,0	
4	14	OT	20,40,40,0,0,0	20, 40: silence frames
14	10	TE	71,72,29,9,0,0	71, 72: burst of [t]
10	10	EE	29,29,0,0,0,0	
10	32	EL	9,207,207,0,0,0	9, 29: parts of [e]
32	1	L-	205,0,0,0,0,0	207, 205: parts of [l]

Table 10. ABU concetanation for synthesis of German hotel

It should be noted that one complete sound combination is totally realized after the processing all the rules, representing the sound combination itself and two immediately adjacent neighbours. For example, when synthesizing the word 'hotel', the speech frames for the [ot] sound combination will only be present in the frame buffer after the HO, OT and TE rules have all been processed. The final version of ABUs for synthesizing the German word hotel are detailed in Table 10 and schematized in Figure 2. The rules of the matrix were manually designed for every sound and sound combination using the interactive development system mentioned above. The final version of every rule was formed also by an interactive refining process of listening, correcting the rule, listening again and so on.

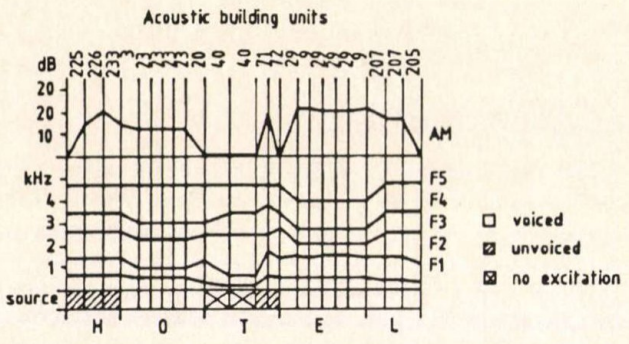


Fig. 2. Schematic ABU specifications of synthetic German hotel

7. STRESS, INTONATION AND RHYTHM

When devising acceptable intonation for unrestricted text, a set of rules has to be formulated which produces natural sounding pitch contours for utterances that may have never been spoken [29]. In the MULTIVOX system, the pitch movement and timing modifications elements of Table 11 are used as modular units in intonation, stress and rhithm generation.

The degree parameter can be applied to all units. Examples: RM means rising to middle level; Ju(SM) means jump up to middle point. The physical values for these three levels are shown in Table 12. These data are valid for male voices.

7.1. Pitch and timing in word stress

Two main factors were taken into consideration in the formation of word stress

- The relation between the pitch variation and the lengthening of accented vowels.
- Vowel duration influences the form of the pitch pattern. Our experience is that the same pitch contour cannot automatically be used in the case of a short and a lengthened vowel. A pitch contour for stressing looks like:

$$\begin{aligned} &\text{for short vowels :} && SPM(RM)(StH)+ \\ & && +(FM)(StM)EPM \\ &\text{for long vowels :} && SPM(RM)(StM)+N(x)+ \\ & && +(FM)(StM)EPM \end{aligned} \qquad (4)$$

The value of (x) is language dependent. For Hungarian and German it is about 30 ms, for Italian and Esperanto in open syllables about 60 ms, and in Finnish about 60 ms.

Element		Code	Associated Specification
Starting point of the pitch contour		S	degree
End point of the pitch contour		E	degree
Direction of the pitch movement:	rise	R	degree
	fall	F	
Steepness of movement in time		St	degree
Jump down		Jd	degree
			S
			E
Jump up		Ju	degree
			S
			E
Lenthening		L	degree
No change		N	ms
Degree:	high	H	
	medium	M	
	low	Lo	

Table 11. Intonation rule elements and their codes

Element	Degree			Unit
	High	Medium	Low	
S/E	125	110	95	Hz
R/F	25	15	5	%
N	-	-	-	-
St	2	0.5	0.25	Hz/ms
L	3x	2x	1.5x	times

Table 12. Values of pitch rule elements

7.1.1. Word stress categories

The five general rules in Table 13 serve for word stress realization in the languages specified.

Number	Rule	Language
1	Stress the first syllable	H, I, G
2	Stress the last syllable	I, G
3	Stress the penultimate syllable	I, E, G
4	Stress other syllables	I, G, D
5	Unstress the sequence	All

Table 13. Stress rules

7.1.2. Stress location in words

Hungarian, Finnish, Spanish and Esperanto can be treated as fixed stress languages, while German, Dutch and Italian are free stressed ones. For fixed stress, the stress location can be determined easily. Afterwards, stress rules are used to superimpose the necessary pitch pattern.

For free stressed languages, the algorithms to find the stressed syllable in the word are based, in many TTS systems, on a large morpheme inventory (10,000–50,000 items) and a morphological analyzer. Such solutions are known for English [2], German [11], [26] and for Italian [14]. The MULTIVOX system was designed to work with a relatively small amount of memory (130-180 kbytes depending on the language) and in real time, even on a slow PC. Therefore, no morpheme inventory is applied at all, instead an indirect approach is used to find the stressed syllable. The process is based, on one hand, on searching for the long vowel in the word (this solves many cases) while on the other hand, special algorhitms are used to find the correct place if the former rule is not valid. (This is used mainly for German; see Table 14.)

Number	Rule	Examples
G1	There is only one stress in one word.	
G2	Stressed prefix or suffix has priority over other rules.	ankommen, Komponist, studieren
G3	An unstressed prefix is followed by a stressed syllable.	bekommen, gesagt
G4	In two syllable words, the long vowel (if any) is stressed.	fahren, sehen, primär
	Otherwise the first is stressed	Silbe, Tausend
G5	Exceptions for special cases are handled with markers in the MT.	

Table 14. Supplementary stress rules for German

Using these rules for finding the place of stress in German words, a correct pitch is superimposed in 95% of the cases. The evaluation of these rules were done by listening to 1600 German sentences [28] and 50 text files (one A4 page each) gathered from books and newspapers. A weaker point of the German word stress assignment is the case of compound words. Here, only rules G2 and G3 can assign a place of the stress for pitch patterns. Incidentally, the correct timing structure (without a pitch pattern superimposed) gives the feeling of correct stressing in most cases.

7.1.3. Unstressing words

The question of unstressing is just as important as stress if a more natural variation is desired among stressed, unstressed, and neutral parts in human speech. Unstressing in MULTIVOX is generated by reducing the pitch value to SL during the sequence (word, prefix, suffix, etc.). In some cases, an amplitude reduction is used as well. These methods are used for every language in the system. In summary, 3 types of patterns are used in word stress generation: stressed, unstressed and neutral sequences, and they remain present in higher level intonation patterns, i.e. in phrase and in sentence intonation.

7.2. Sentence intonation

In sentence intonation, one serious problem is to find such rules that make the monotonous speech more natural, so that listening to long texts would not be uncomfortable. Declarative and questions intonation contours are generated automatically in the MULTIVOX system.

7.2.1. Declarative sentence

For declaratives, the general theoretical pattern is a linear falling one. This pattern is used for all the languages excpet Italian, where a rising-falling pattern is superimposed. The final falling intonation curve will be constructed considering three types of parameters i.e. the length of the sentence (determined by the word counter), the starting value of the fundamental frequency and the intonation of the last word of the sentence. The subcases and the characteristic changes are shown in Table 15. At phrase bundaries, the pitch is set higher (2–4 Hz per boundary) in the last two categories. This gives the feeling that a new phrase has begun. With these simple rules, a relatively diversified intonation has been reached when reading long texts.

Words in the sentence	Starting pitch level	Falling intonation rule		Last word pitch drop
1-2	-		PI=10 x -2 Hz	-
3	-2 Hz	in word 1,2	PI=-12 Hz	PI=- 15 Hz
4-5	-4 Hz	in word 2, 3	PI=-10 Hz	PI=- 12 Hz
6-10	-6 Hz	in every word	PI=-2 Hz	PI=- 10 Hz
11-20	-7 Hz		PI=10 x -1 Hz	PI=- 9 Hz
21-	-8 Hz		PI=12 x -1 Hz	PI= -8 Hz

Table 15. Pitch settings for declarative sentence intonation

7.2.2. Questions

In case of questions, different types of pitch patterns have to be superimposed depending on the kind of question, like questions with a Q word, and without a Q word and one-syllable questions.

7.3. Rhythm processing

The timing scales are different between a written text and its spoken version. In the written form, the unit is the word, while in the uttered form, it is the so called 'prosodic word'. This means that, in natural pronounciation, the words are not pronounced separately one after the other (as they are written), but instead prosodic words (phrases) are pronounced fluently as one unit and pauses are kept between them. The better rules are created to find the boundaries of the prosodic words, the closer the prosody of the machine voice will be to the natural one.

7.3.1. Handling comma like effects

The acoustic realization of comma, colon or semicolon is performed by keeping a pause and superimposing an intonation contour which is characteristic of a comma.

7.3.2. Phrase structure determination

In order to detect the phrase structure in a text, a complex grammatic and semantic analysis would be necessary [2], [11]. This requires a large amount of computational effort and it is time consuming. The problem is even more complicated in a multilingual structure. Therefore, compromises had to be made and reliable solutions had to be developed which can solve the problem to a relatively

high degree without syntactic and semantic analysis. In the MULTIVOX system both language-independent rules and specific language-oriented rules serve to determine the location of tha future breaks in the text.

An example for a forward joining general rule for rhythm generation is given in Table 16. The forward joining rule: articles, personal and possessive pronouns and other function words (like prepositions) are joined together with each other and with the following word; also question words are joined to the next word.

Written form	Spoken form	Language	Gloss
In meinem Haus ist...	InmeinemHaus ist	G	In my house there is...
Sur la tablo estas	Surlatablo estas	E	On the table there is...
La mia mama viva in Roma	Lamiamama viva inRoma	I	My mother lives in Rome
Az asztal nagy	Azasztal nagy	H	The table is big
Was will Paul von mir?	Waswill Paul vonmir?	G	What does Paul want me to do?
Kio estas sur la tablo?	Kioestas surlatablo?	E	What is on the table?

Table 16. Prosody processing

8. EVALUATION OF SPEECH QUALITY

Two types of measurements were carried out to evaluate the quality of automatically generated speech. The acoustic quality of speech sounds was tested by listening to one and two syllable words, and the general speech quality was tested by listening to short stories and news (5–10 sentences) automatically read by the system from text files. Scientific evaluation tests were carried out for Hungarian twice during the development process [9]. About seventy 18–40 year old subjects took part in tests. The final intelligibility results are: at the word level 85%, at the sentence level 95%

As for naturalness, the subjects judged the voice as not very robot-like. For the other languages, listening tests were carried out individually with 3–6 native people and the system was demonstrated in several exhibitions and conferences like: Speech research '89 in Budapest; Hannover Industry Fair 1990 [3], Rome 1990, Autrans 1990; IPO Eindhoven, Amsterdam, Groningen 1990; Expo-Lingua Wien, 12th ICPhS in Aix-en-Provence, Eurospeech conference in Genova 1991. MULTIVOX was one of the participants of the finals of the Software for Europe competition held during the Hannover CeBIT'92 exhibition and fair.

The general opinion of the listeners was that the voice quality is good and the automatically generated speech is inlelligibile.

REFERENCES

[1] Ainsworth, W. A., "A system for converting English text into speech", IEEE Transactions on Audio and Electroacoustics, vol. AU–21 1973, pp. 288–29.

[2] Allen, J., Hunnicutt, S., and Klatt, D., *From text to speech: The MITalk system*, Cambridge University Press.

[3] Burgwitz, A., "Lesen und lesen lassen. PC Sprachausgabe MULTIVOX", *Magazin für Computer Technik*, 1991, pp. 102–103.

[4] Carlson, R., Gramnstrom, B., and Hunnicutt, S., "A multilingual text-to-speech module", *IEEE Conference on Acoustics, Speech and Signal Processing*, 1982, pp. 1604–1607.

[5] Coker, C. H., Church, K. W., and Lieberman, M. Y., "Morphology and rhytming: Two powerful alternatives to letter-to-sound rules for speech synthesis", *Proceedings of the ESCA Workshop on Speech Synthesis, Autrans*, 1990, pp. 83–87.

[6] Delmonte, R., Mian, G. A., and Tisato, G., "A text to speech system for Italian", *IEEE Conference on Acoustics, Speech and Signal Processing*, vol. 3, 2.9, 1984, pp. 1–4.

[7] Freenkenberger, S., Kommenda, M., and Wirth, Z., "Halbsilben Inventar der deutschen Sprache für einen Formantsynthetisator", *Fortshritte der Akustik, DAGA '91, (Gemeinschaffstagung der Deutschen Arbeitsgemeinshaft für Akustik*, Part B, 1991, pp. 953–956.

[8] Fujimura, O., and Lovins, J., "Syllables as concatenative phonetic elements", *Syllables and segment (A. Bell and J. B. Hooper, ads.)*, Amsterdam, North Holland, 1987, pp.107–120.

[9] Gosy, M., Laczko, M., and Olaszy. G., "The evaluation of speech quality of the MULTIVOX text-to-speech system", (in Hungarian), *Hungarian Papers in Phonetics*, vol. 23, 1991.

[10] Kiss, G., Olaszy, G., "An interactive speech synthesizer system with computer and OVEIII speech synthesizer" (in Hungarian), *Hungarian Papers in Phonetics*, vol. 10, 1982, pp. 21–47.

[11] Kohler, K., "Improving the prosody in German text-to-speech output", *Proceedings of the ESCA Workshop on Speech Synthesis*, Autrans, 1990, pp. 189–192.

[12] Koutny, I., "Speech synthesis and its application in teaching Hungarian and Esperanto", (in Hungarian), *Ph. D. Dissertation.* Technical University of Budapest, 1990.

[13] Marty, F., and Hart, R. S., "Computer program to transcribe French text into speech: Problems and suggested solutions", *Technical Report of the University of Illionis, LLL-T-6*, University of Illionis, Urbana, 1985.

[14] Martin, P., "Automatic assignment of lexical stress in Italian", *Proceedings of the ESCA Workshop on Speech Synthesis*, Autrans, 1990, pp. 149–152.

[15] Nemeth, G., Gordos, G., and Olaszy, G., "Implementational aspects and the development system of the MULTIVOX text-to-speech converter", *Proceedings of the ESCA Workshop on Speech Synthesis*, Austrans, 1990, pp. 233–237.

[16] Olaszy, G., "Preparation of Computer Formant Synthesis of Sound Sequences", (in Hungarian), *Hungarian Papers in Phonetics*, vol. 8, 1981, pp. 147–160.

[17] Olaszy, G., "Some rules for the formant synthesis of Hungarian", *Proceedings of the 8th Acoustic Colloquium*, Budapest, 1982, pp. 204–210.

[18] Olaszy, G., "Speech synthesis in Hungary from the beginnings up to 1989", *Proceedings of the Speech Research '89 International Conference*, Budapest, 1989 , pp. 289–292.

[19] Olaszy, G., "Electronic speech generation: The acoustic structure and formant synthesis of Hungarian", (in Hungarian), Budapest, 1989.

[20] Olaszy, G., "MULTIVOX-A flexibile text-to-speech system for Hungarian, Finnish, German, Esperanto, Italian and other languages for IBM PC", *Proceedings of the European Conference on Speech Communication and Technology*, vol. 2, 1989c, pp.525–529.

[21] Olaszy, G. and Gordos, G., "On the speaking module of an automatic reading machine", *Proceedings of the 11th International Congress of Phonetic Sciences*, Tallin, vol. 3, 1987, pp. 93–97.

[22] Olaszy, G., Gordos, G., and, Nemeth, G., "Phonetic aspects of the MULTIVOX text-to-speech system", *Proceedings of the ESCA Workshop on Speech Synthesis*, Autrans, 1990, pp. 277–280.

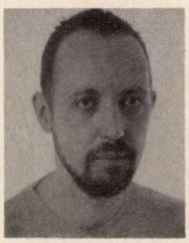
[23] Olive, J. P., "A new algorithm for a concatenatier speech synthesys system using an augmented acoustic inventory of speech sounds", *Proceedings of the ESCA Workshop on Speech Synthesis,*, Autrans, 1990, pp. 25–29.

[24] O'Shaughnessy, D, "Parsing with a small dictionary for applications such as text-to-speech", *Computational Linguistics*, vol. 15-2, 1989, pp. 97–108.

[25] Peterson, G., Wang, W., and Silversten, E., "Segmentation techniques in speech synthesis", *Journal of the Acoustical*

Society of America, vol. 30, 1985, pp. 739–742.
[26] Pounder, A., Kommenda, M., "Morphological analysis for German text-to-speech synthesis", *Proceedings of the 11th International Conference on Computer Linguistics*, Bonn.
[27] Salaza, P. L., "Phonetic transcription rules for text-to-speech synthesis of Italian", *Phonetica*, vol. 47, 1990, pp. 66–83.
[28] Sotschek, J., "Sätze für Sprachgütemessungen und ihre phonologische Anpassung an die Deutsche", *Fortschritte der Akustik, DAGA '84 (Gemeinschaftstagung der Deutschen Arbeitsgemeinschaft für Akustik)*, 1984, pp. 873–876.
[29] Terken, J., and Collier, R., "Designing algorithms for intonation in synthetic speech", *Proceedings of the ESCA Workshop on Speech Synthesis*, Autrans, 1990, pp. 205–208.
[30] Wotke, K., "From ortography to phonetic transcription in the German text-to-speech system TETOS", *Proceedings of the ESCA Workshop on Speech Synthesis*, Autrans, pp. 219–222.

Gábor Olaszy graduated at the Technical University of Budapest in 1967. He has been with the Phonetics Laboratory of the Linguistic Institute of the Hungarian Academy of Sciences since 1974. In 1986 he was appointed head of this laboratory. Main fields of his activity are (i) the research of the acoustic structure of speech, (ii) developing formant-based speech synthesis (iii), text-to-speech synthesis (iv) embedding speech synthesis technics into applications (language learning, medical usage, solutions for the handicapped etc.). In 1985 he received the dr. tech. degree at the Technical University of Budapest, and two years later, in 1987 the Ph.D. degree at the Eötvös Lóránt University, Budapest. He has 70 publications (including two books) in different languages. He is co-author of 5 patents concerning new processes in speech synthesis technics and also application of synthesized speech in practice.



Géza Németh graduated in 1983 at the Technical University of Budapest, Faculty of Telecommunication Engineering. He was granted a scholarship in digital signal processing, and received his Ph.D. degree in 1987 for his work on formant synthesis. In the period between 1985 and 1987 he has been working at BEAG on the development and application of speech synthesis and speech recognition systems. Since 1987, he is with the Department of Telecommunications and Telematics, Technical University of Budapest, now as assistant professor. His field of interest covers speech technology and lately telecommunication managment networks.

FULL HUNGARIAN REAL-TIME TEXT-TO-SPEECH SPECIALLY DEVELOPED FOR THE BLIND

A. ARATÓ

KFKI INSTITUTE FOR MEASUREMENT AND COMPUTING TECHNICS

Since 1985 more than 500 BraiLab talking computers have been used by blind pupils, students, individuals and workers in Hungary. In this paper we summarize human engineering problems of user friendly interfaces, based on a synthetic speech output computer-aid for the blind. It is emphasized that during the development of artificial speech the requirements of the visually handicapped have to be always taken into account.

1. INTRODUCTION

Our first attempt to create aid for the blind in the 80's was a one-cell "soft" braille output display integrated in a dumb terminal. The pins were driven by solenoids, and the only way to check the screen was off-line capability. After that we developed a talking System Design Kit for Intel 8085 based on the Digitalker with fixed vocabulary. Soon we recognized that the most appropriate solution for the blind user was the full Hungarian real-time text-to-speech based intelligent computer aid. Our first talking computers were built with MEA-8000 formant synthesizer and Z80 microprocessor. In an earlier paper we reported on BraiLab and BraiLab Plus Hungarian and German talking computers [1]. (The BraiLab home computer with Basic interpreter was the first talking personal computer in Hungary.) These machines were supplied with a very compact full text-to-speech program due to the method of storing speech parameters and Metabraille coding [2]. The disadvantages of this text-to-speech program were noisy speech and poor sound because of few speech frames. In the third member of our BraiLab family, BraiLab PC (for IBM PC), these problems have been eliminated.

2. COMPACT REAL-TIME TEXT-TO-SPEECH IN BRAILABS WITH MEA-8000

In 1984 there was not available a microprocessor based full Hungarian text-to-speech which could have been used for a computer-aid for the blind persons. The Linguistic Institute of the Academy of Sciences developed a quasi real-time, very good quality Hungarian unlimited vocabulary synthetic speech. The program was written in FORTRAN with overlay technique for a PDP-11 minicomputer. It was generally thought, that a small microcomputer had not enough capacity to realize full real-time text-to-speech and other ordinary tasks at the same time. In spite of this general opinion we started to develop special synthetic speech for the blind.

The requirements for synthetic speech in a computer-aid for visually handicapped are quick response, accuracy and efficiency. The good quality of the speech is important mostly for reading machines. Our main goal was to create a very cheap, but efficient talking home computer for visually impaired persons in Hungary. Our first talking System Design Kit (SDK85) for Inter 8085 served as a development system to create the text-to-speech program and the data base for MEA-8000 formant synthesizer.

We introduced an overlapping technique to store the speech frames of 33*33 dyads of the Hungarian language. Due to this storing method, our speech data base required only 4 Kbyte of memory including the dyad matrix. This technique required a special coupled computer development system. In the coupled system the previously mentioned talking SDK85 was connected to the TPA/1148 (KFKI PDP-11 compatible) computer (Fig. 1). The latter computer held the speech frame compiler, which translated the source labels and speech parameters to the matrix and the overlapped data base.

The function of the talking SDK85 was to develop the program itself and to generate speech with new data base. Digitalker helped my wife Therese Vaspöri, who is totally blind and took part in the developing process. She typed in sample text with our one hand braille keyboard of the SDK85. The SDK85 was completed with more memory, serial port and of course also with MEA-8000 formant synthesizer.

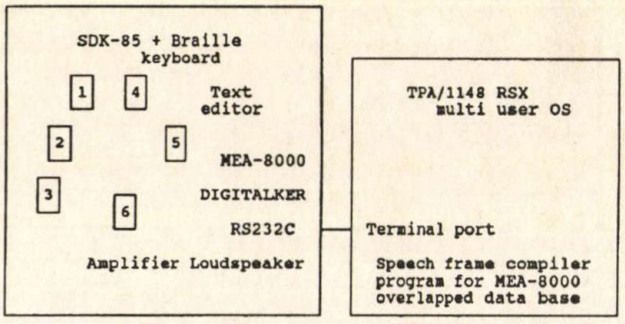


Fig. 1. Development System to Create Data Base for MEA-8000

The braille input has a special importance in our system not only because of the special application, but also due to the inner structure of our text-to-speech program. For speeding up the ASCII to phoneme translation and facilitating braille publishing, we intorduced Metabraille coding. The six dot braille position has the structure shown in Fig. 2.

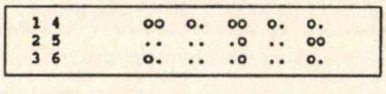


Fig. 2. The layout of the six braille dots

The sample word in Fig. 2 is "magyar" (Hungarian). The sound "gy" which is written in Hungarian normal writing with two letters, in braille is written with one cell.

The Hungarian (and the German) uncontracted braille is really contracted. These "abbreviations" are nearly phonemes as it is shown in Fig. 3 and Fig. 4.

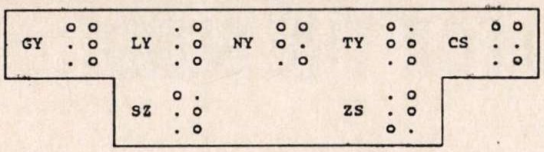


Fig. 3. Hungarian Braille Contractions

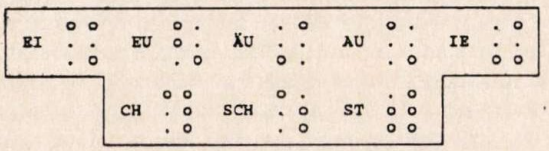


Fig. 4. German Braille Contractions

So if we wanted to write the word "magyar" in metabraille, it would look "maGar". The metabraille coding gives great advantage not only in creating compact inner exception dictionary of text-to-speech, but especially in braille publishing system. The braille cell has one-to-one correspondence on the screen, so blind and sighted people can easily edit braille texts with word processors.

We built our Hungarian text-to-speech system into two Z80 based machines. The first machine was called BraiLab Basic. This talking aid for the blind primarily served for school and home purposes. From 1985 about 400 devices have been used by Hungarian blind users. BraiLab has only talking BASIC interpreter and a simple talking assembler-monitor. BraiLab Basic became the official school computer of the Blinds' Primary School. Teachers and blind persons themselves wrote a lot of programs for many subjects and games.

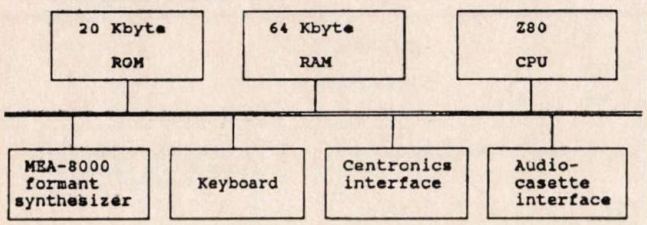


Fig. 5. Configuration of BraiLab Basic

Our second aid called BraiLab Plus is aimed for rehabilitation prupose. Many visually handicapped persons could find their job due to BraiLab Plus. There are 60 BraiLab Plus computers used by blind users. These computers provide talking Wordstar, talking data base manager. The operating system is integrated with speech output.

The keyboard of BraiLab Plus can be used not only as normal QWERTY, but as braille keyboard too. The metabraille coding allows easy interactive braille text editing. The Hungarian braille has grad 1 ("uncontracted") and two versions of grade 2 (44 or 77 contractions) writing. For all these versions we wrote an ASCII to metabraille compiler. Later we used the exception dictionary handling rutine of this compiler in our next text-to-speech system for PCF-8200. The configuration of BraiLab Plus computer is shown in Fig. 6.

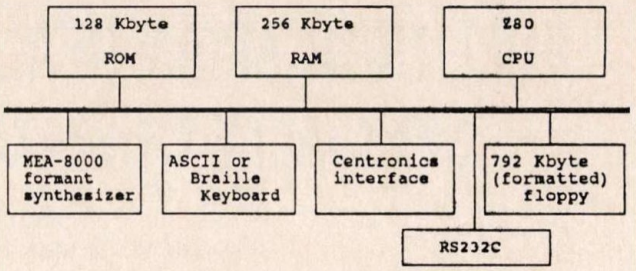


Fig. 6. Configuration of BraiLab Plus

3. SPEECH UNDERSTANDABILITY AND NATURALITY IN BRAILAB PC

In our opinion, computer-aides for the blind require different speech output than ordinary applications for sighted persons.

In judging artificial speech, two characteristics play main roles: understandability and naturality. In the case of a computer-aid for visually impaired persons understandability is more important. Unfortunately satisfying these two demands are often contradictory. It is difficult to solve this confusion.

Firstly let us study the need of understandability. It seems to be unnecessary to deal with meaningless sounds (logatoms), however this is not the case. Think of the fact that speech is often the only help during word processing.

In our case we would like to introduce the term: endurance to the artificial speech instead of naturality. When listening to long texts the naturality of intonation and accent helps the understanding. At present it is not possible to perform real-time semantic analysis of text with the capacity of available microprocessors.

Further we will take examples from our BraiLab PC artificial speech system to demonstrate these two contaidicting but sometimes complementing properties. The experiences with numerous BraiLabs have led us to create a new computer-aid for the blind using IBM PC, with new text-to-speech data base and program (Fig. 7).

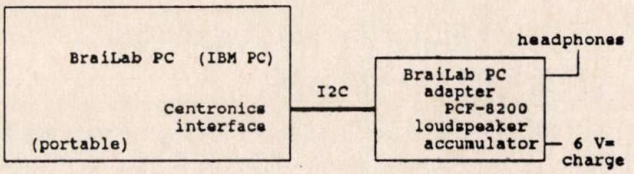


Fig. 7. Configuration of Portable BraiLab PC

First let us consider the problem of the speed of speech. The chosen synthesizer (PCF-8200) strongly limited our realization. We could have prepared longer frames for dyads, however, the shortest ones were chosen. Controlling the frames speed to the minimum, we could have set half frame duration achieving with it faster tempo. Although tempo is one of the most important human engineering demands, we did not build our speech from frames longer than standard duration. It was a hard decision, but we preferred rich voice with more details and "micro-intonation" to faster tempo but less endurance.

We are satisfied with the four different frame speeds of the chip. We did not use frame speed to express

accent. These four speeds are used for user settings. If we had used frame speeds to express accent, we could have advanced the naturality of the speech. In Hungarian Linguistic Institute Gábor Olaszy researched the micro-intonation structures of languages [3]. He realized micro-intonation by changing the pitch. We did not follow this way of achievcing some kind of rhythm. We stored instead in the formant structure of the speech some kind of micro-changes. It is clear, that judgement of the results (naturality endurance) is rather subjective and depends on how well somebody is accustomed to the speech.

After dealing with problems of tempo and "micro-intonation", let us consider the difficulties of understanding sounds and logatoms. One can ask: why is it critical to understand smaller parts of speech than words? During text editing, meaningless components of words can be heard.

For key echo we introduced the sound forms of letters.

From a human engineering point of view key echo is a corner stone of our research. In Hungary we used firstly the echoing with sound forms. You can often find aids for the blind with letter name echo of keys, even in phonetic like languages.

It is an erroneous solution because the user is hindered in typing. The typist is forced to wait till the end of the name of the letter. Even if the speech can not be interrupted this type of echo may make the user nervous. When echoing the words only at word delimiters, you loose the advantage — the good feeling of continuous real-time control. In some languages, like English, this delayed echo seems to be the only solution.

Let us consider the question of resounding Hungarian l;etters eL, eM, eN and eF with their names as shown. if the user types adequately fast, and the speech can be interrupted, than the only sound E will be caught on. It is fully misinterpreted. Now this conclusion appears trivial, but some computer-aides failed in Hungary which did not use proper key echo.

The question has not been answered yet: which kind of speech frames must sound in text-to-speech program during key echo? One possible choice is fitting space-sound dyad with sound-space dyad. This choice has to be considered in speech parameter development. We are compromising with nautrality. The other attempt is the storage of additional parameters with the purpose of sound form echo. It necessitates more memory for implementing in BraiLab PC, therefore we rejected it. For future use, the program code has been realized.

Differentiating upper and lower case is also an important task. Think about C language programming. Braille writing often neglects upper case letters, however our experience has shown the importance of indicating upper case letters. We introduced a new natural signaling of upper case, which does not require additional time, but nevertheless is very attractive. Our capitals sound louder. This displaying method seems to be very simple, but it was rather hard to realize in our earlier devices. Speech frames were prepared with such amplitude parameters, that even with loud vowels we could produce louder sound for capitals. This is the reason, that our synthetic speech produced by MEA-8000 appears slightly noisy. Fortunately in case of PCF-8200, using DAC amplitude factor we could eliminate this noise problem [4].

The importance of logatom and word understandability will be clearer if we explain our echo system of text editing. During word processing correcting letters and words, you can hear only parts of sounds and words concatenating with correcting sounds, resulting at last in the right word pronounced. The user doesn't know consciosly all details of this sound–text editing he or she only uses standard editor commands.

After discussing word understandability we have reached the problems of endurance listening to long texts. For this the sounds have to be stored in a form being rich in details so the question arises how can we store formants with rich details using only 20 Kbytes of speech frames? The answer is in our coding technique. We consider whole section of frames and not single frame as base unit. One dyad can consist of more than 10 frames. The sections can be overlapped, so the whole program including the screen reader is less than 64 Kbytes mainly written in assembly language.

Developing automatic intonation for text-to-speech, we were thinking of blind users. The macro-intonation applied by us indicates the end punctuations already at the beginning of macro-intonationed sentences. We use 6-7 types of macro-intonation helping users not to become tired of long texts.

To improve understandability we recorded articulated speech for every sound and not natural half-pronounced one. This was our compromise towards understandability. We have a special switch in text-to-speech system to increase understandability of text by separating sounds.

4. OVERLAPPED SPEECH DATA BASES

There is a method of storing speech frames by enumerating them, and storing the numbers of the frames for a given dyad in the dyad matrix itself. Limiting the code number to one byte, only 256 different frames, can be used for producing synthetic speech. Increasing the possible number of frames, the matrix will grow considerably. The method has the disadvantage that even using few frames can not reduce the size of the matrix because of its fix pointer structure.

We introduced a new technique in storing speech frame prototypes. We store only a count and a pointer in dyad matrix, so we can use variable length frame sections. Grouping the different parts of these sections we can overlap them. This technique is very useful when we have small data base (in case of MEA-8000 4 Kbyte together with the matrix!). The technique is also advantageous for large data base with many details of speech, when frames will have some micro-changes in bandwidth and formant parameters. Using this overlapped technique, in case of PCF-8200, the size of the data base is reduced to 20 Kbyte (with dyad matrix).

Regarding MEA-8000's data base, in consonant-vowel dyads we stored only frames characterizing transitions. Often these frames contain zero amplitude, allowing MEA-8000 to do linear interpolation of the parameters. The consonants are stored in vowel-consonant dyads. Of course this method leads to a rather rough but well understandable speech. Several examples of dyads prepared for MEA-8000 in BraiLab Basic and BraiLab Plus will follow. The numbers are written in order as it is shown in Fig. 8.

```
FD AMPL  PI    F1   B1   F2   B2   F3   B3   B4
```

FD=frame duration msec; AMPL=amplitude; PI=pitch increment; F1, F2, F3 = formant frequencies; B1, B2, B3, B4 = formant bandwidthes, given in Hz. PI=16=noise.

Fig. 8. The Header for MEA-8000 data base

In the case of PCF-8200, we stored two section pointers with appropriate frame counters in the dyad matrix. So we could generate speech very rich in details, however with overlapping sections we were able to reduce the size of data base. One dyad can consist of 15 frames.

In the consonant-vowel dyads we stored the frames characterizing the given consonant (with non zero amplitude), the transition to vowel and the vowel itself. In vowel-consonant dyads we stored only the frames of transtiion and frames with zero amplitude (if such frames existed).

All these frames have minimal frame duration, so we could store micro-changes in formant frequencies and bandwiuth parameters. Next figure shows the header for PCF-8200 data base. Here the numbers are given as codes, and not as msec or Hz values.

```
FD AMPL      PI    F1 B1  F2 B2  F3 B3  F4 B4  F5 B5
```

FD=frame duration msec; AMPL=amplitude; PI=pitch increment; F1, F2, F3, F3, F5 = formant frequencies; B1, B2, B3, B4, B5 = formant bandwidthes, given in Hz. PI=16=noise.

Fig. 9. The Header for PCF-8200 data base

In Fig. 10. and Fig. 11. you can compare the overlapped technique used for small compact data base (MEA-8000) with overlapped rich data base (PCF-8200) which was used in BraiLab PC.

5. MAN–MACHINE RELATIONS WITH ARTIFICIAL SPEECH

There is an interesting psychological inhibition in users, when a computer speaks. At the beginning everybody waits till the end of the pronounced message and does not touch any key. (Sighted persons do not like to use speech output, if only the same message is produced as on the screen, because he or she can read faster from visual display than listening to confusing voice). After training, blind persons require faster working rate then the quickest tempo of the synthesizer. The most important man–machine requirement is to fulfill fastest working demand of the blind.

The artificial speech is the fastest available interactive output for the blind. A speaking aid is usable only if it conveys the most important part of the screen's information and it conveys only that one. This goal can be accessed fully and easily, with specially developed devices and programs. In this case, however, visually impaired persons are excluded from numerous hardware and software products developed for sighted people, but useful for the handicapped too. Our Basic BraiLab and BraiLab Plus computers represent a special aid, but BraiLab PC belongs to the latter category.

Special assistive devices can be more fool-proof than general ones. Fool-proofness can be improved with an

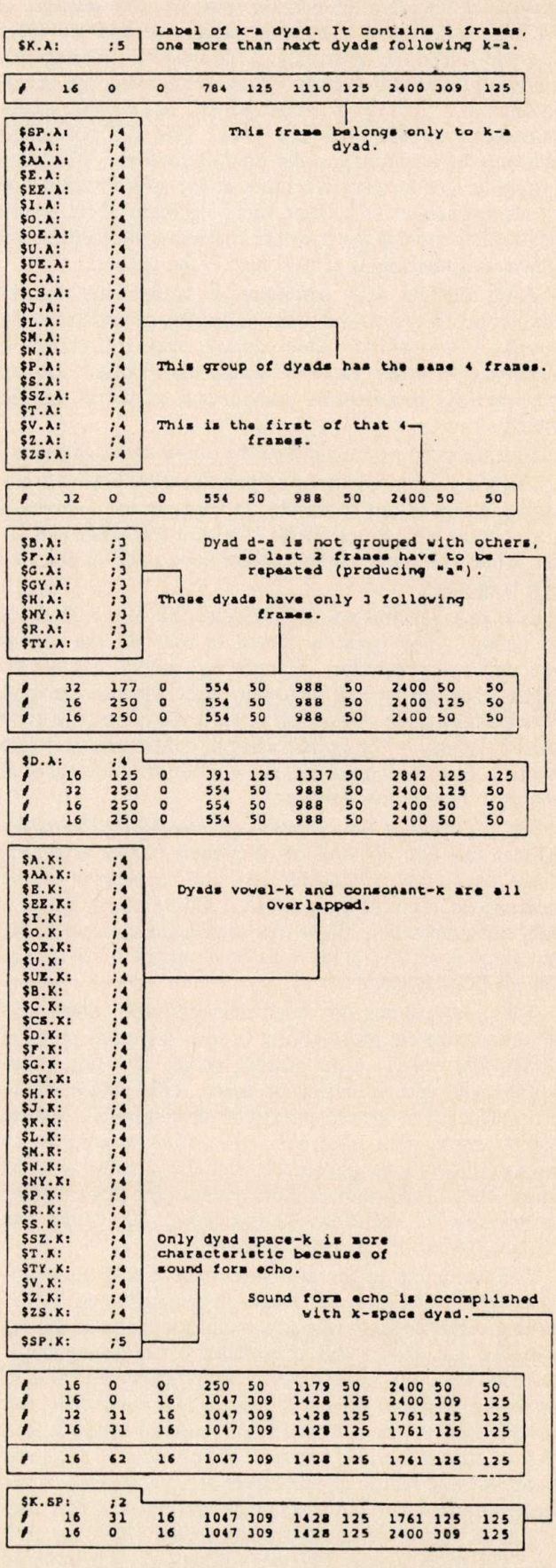


Fig. 10. Overlapped data base for MEA-8000

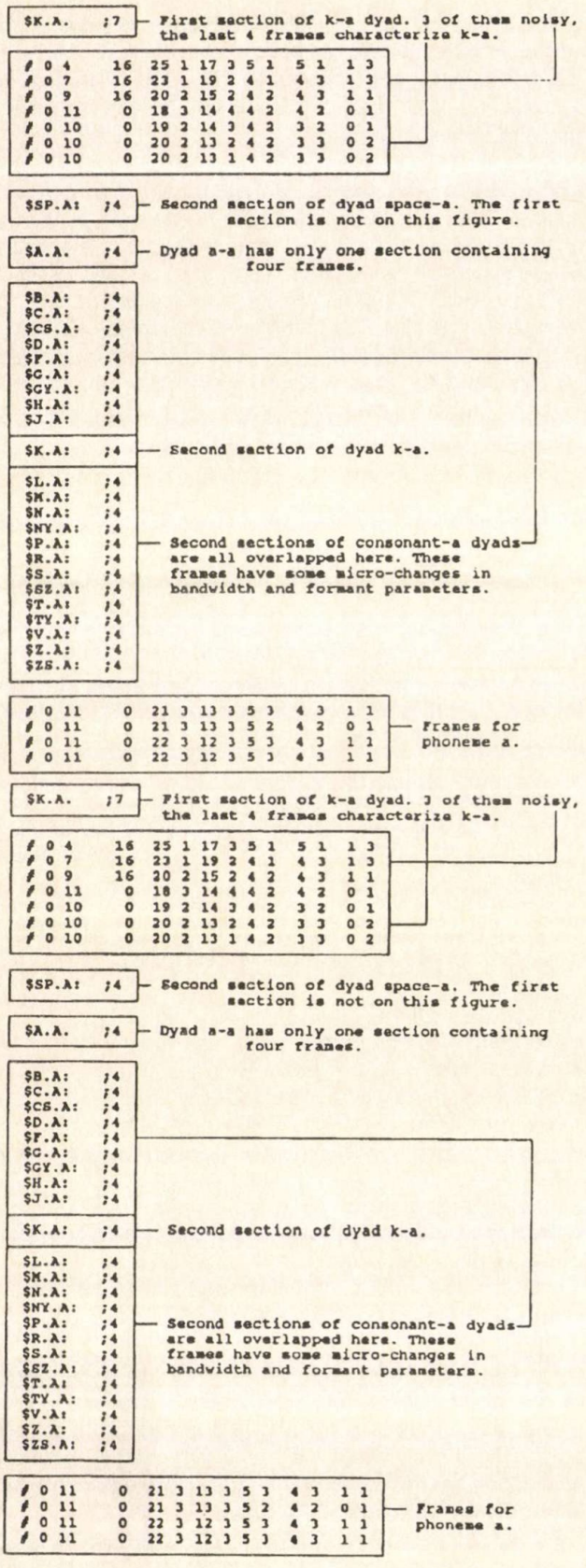


Fig. 11. Overlapped data base for PCF-8200

intelligent screen reader system. When can one call a screen reader intelligent? This question will not be answered fully by us, but we would like to show some of our basic solutions in this paper.

Our aim was that blind users should not have to learn a screen reader command set. It is enough for them to practice only with the control of a given program. An intelligent screen reader has to "know" when and what should be outputted to the synthesizer.

Therefore we integrated the screen reading and keyboard echo functions in our text-to-speech system. Using basic editing functios of word processors, one need not have to deal with echoing, it is automatic. By manipulating on-line in the text with lines, words and letters they are outputted to the formant synthesizer properly. The correctness of the letter, word or line can be heard at once.

Our screen reader handles windows on the display. Windows as well as graphics and colors, are disturbing for blind persons.

In many computer-aides you can hear special sound effects. We preferred special synthesized voice effects to beeps. Color or inverse attributes sound can be distinguished from normal ones.

6. CONCLUSIONS

An ordinary user never feels consciously the man–machine problems. The blind person has only an impression about the user friendliness of the aid. The assistive device seems comfortable if you do not need to pay attention to the use of the aid very much, but you can concentrate on the real job. That's why you have to continue special speech technology research on technical aids for the handicapped.

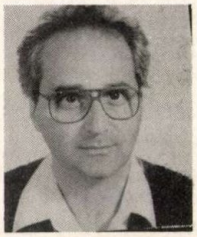
REFERENCES

[1] Arató, A., Vaspöri, T., Olaszy, G., "Hungarian and German Speaking Computers for the Blind". Beyond Number Crunching Austrian–Hungarian Conference OCG-NJSZT, 1988.

[2] Arató, A., Molnár, P., Vaspöri, T., "Computer-Aided Hungarian Contracted Braille" 2nd ICCP OCG, 1990.

[3] Olaszy, G., "Electronic Speech Producing" (in Hungarian), Műszaki Könyvkiadó, Budapest, 1989.

[4] Zelle, H.W., ten Have, M., "Appliation Report for the PCF-8200 Formant Speech Synthesizer". Philips Report No. EDP-8807.



András Arató obtained his M.Sc. in electrical engineering in 1974 from the Popov Electrical Institute of Saint Petersburg. He is working in the KFKI Institute for Measurement and Computing Techniques. His previous fields of activity were high speed computer communications, local area networks, microprogramming. In the last ten years he has been involved in the special speech technology research for rehabilitation of handicapped persons. He developed the speaking BraiLab computer family for the blind. He is also teaching visually handicapped students at the Eötvös Loránd University, Budapest.

SOME HUNGARIAN PRODUCTS AND SERVICES IN SPEECH PROCESSING

Talking and listening are two fundamental activities. Machines that "talk" and "listen" can extend human voice communication in many useful ways. Voice processing equipment does just that by providing an audio information channel and a unique human-computer interface.

Voice processing technology is the result of developments in computers, digital signal processing, large scale integrated circuits and software. Despite its foundation in high technology , voice processing has a lot to do with people:

- helps people communicate with other people more efficiently,
- helps people conveniently access recorded information and computer databases,
- helps people carry out a variety of computer-based tasks,
- helps people control and communicate with many kinds of devices (such as telephones or appliances) that provide non-computer functions.

People, in fact are the most important part of any voice processing installation. The applications described in this paper illustrate the practical results of putting voice technology to work for people. The following products and services were developed at the Department of Telecommunications and Telematics, Technical University Budapest (DTT,TUB).

SPEECH SYNTHESIS: THE MULTIVOX MULTILINGUAL TEXT-TO-SPEECH SYSTEM

Text-to-speech (TTS) systems are capable of converting unlimited vocabulary written text to speech. Many people think these are for visually impaired and blind users only. The reason for this may vary from the poor quality of early TTS solutions to the fact that application designers are used to conventional input/output devices (i.e. keyboard, the screen and lately the mouse).

The latest TTS systems already offer a wide range in both performance and price. This provides a way for application designers and users to apply natural means of communication: speech. Applications are not bound to the screen or user-unfriendly beeps anymore. Any output information can be made audible and intelligible, multimodal dialogues can be created.

This can sometimes virtually save lives. In case of the largest nuclear accident of the Western hemisphere, the Three Miles Island power plant in the U.S.A., one of the operators threw a newspaper on the control desk. This prevented at the critical time to recognize the flashing alarm light. This could not have been the case with voice message.

A less critical, but still rather boring experience is that of those using graphics program in PC network. Network alarm messages (e.g. call for lunch) quite often can not get through. A TTS system coupled to proper application programs could help.

MULTIVOX, described in a paper of this issue, is a general purpose, multilingual, programmable text-to-speech system. The open architecture of the system allows the use of the speech output facility in any application. Engineers, programmers are thus offered an effective tool to teach their software to speak. The small memory requirement allows the usage of this speech system in parallel with other programs, giving them the most effective human information transmission method, the use of speech. The multivox software family ensures the usage of this text-to-speech system for non-expert computer users.

MULTIVOX speaks (at present) eight languages: standard modern Arabic, Dutch, Esperanto, Finnish, German, Hungarian, Italian and Spanish. New languages under development are French and English. This feature ensures that applications, using MULTIVOX for one language can easily be ported to many European and Middle-East countries, if needs arise. (This can save development time, cost and open new markets for application developers, and may result in lower cost for end-users.) Developers are continuously working on development toward better speech quality (**MULTIVOX-2**).

Basic feature of MULTIVOX are:

- input: unlimited text input with normal orthography
- output: good quality speech with proper intonation and rhythm
- phonetically based formant synthesis
- use of finite set of Acoustic Building Units (ABUs)
- highly structured grapheme-to-sound and sound-to-ABU conversion
- simple installation
- small memory requirement
- resident text-to-speech program
- polling or interrupt driven operation
- real-time operation from PC/XT to AT/486
- programmable speech options (speed, volume, pitch, intonation etc.)
- two voice types
- user extendable exception vocabulary for foreign words, names etc.
- built in abbreviation recognition and pronounciation
- operation systems: DOS, Windows, OS2

MULTIVOX means additionally a broad software family. Members of this family are the followings.

GRAPHVOX is a simple word processor with speech output facilities. Keyboard operations, cursor position, text editing and file operations (load, save, directory) can be supported with spoken messages.

PAROLERN is a speaking language learning program package. Dictation, auditive comprehension, spelling and grammatical excercises, etc. are included. Individual courseware can be freely created.

PHONOVOX is an interactive speech editing program for scientific and teaching purposes. A special screen editor displays speech frame data of the synthesizer (formants, bandwiths etc.) which can be edited at will. The

user has complete freedom as to change speech data. Immediate speech echo of the edited speech signal serve effective research and demonstration.

READTEXT is a text file reader.

SAY utters the sentence typed on the keyboard.

VORTAR is interactive vocabulary in which the items of the target language can be listened to as well.

VOXAID is a special communication aid program for speech impaired (this works in the portable PORTalker device, which is actually a laptop XT with the MULTIVOX hardware integrated into it).

VOX-CONTROLL is a utility to connect a PC based text-to-speech output to any other computer type with an RS-232 asynchronous port.

MULTIVOX has also been licensed by the Austrian Research Center at Seibersdorf who use it in a cheap, AMIGA 500 based reading machine for the blind.

MULTIVOX system components

- MULTIVOX hardware device
- power supply (220V/12V DC, 300mA)
- headphones
- floppy diskette.

System requirements

- IBM XT, AT, AT/386, AT/486, PS/1, PS/2 computer or compatible (at least 16MHz AT recommended) with minimum 512KByte RAM (DOS), for Windows or OS/2 as required by the supplier.
- DOS 3.x or later, Windows 3.0 or later and OS/2 1.2 or later.
- one parallel (Centronics) printer port
- one floppy drive
- IBM MDA, CGA, EGA, VGA or Hercules card and monitor.

PORTalker is a portable small size (notebook) speaking aid for speech impaired persons. This device can be used effectively in hospitals (for patients after larynx operations for people suffering in certain aphasia etc.) and in everyday life as well. PORTalker has typed text input (or called presaved sequences from the memory on the keyboard), and provides a good quality speech output with correct intonation and rhythm.

The services of the PORTalker speaking aid are as follows:

- telling unlimited (in length and content) texts after typing in,
- flexible editing and presaving of 100 messages statements and calling any of them to be spoken by pushing only one button,
- presaving for constant use 50 frequently used statements (especially in hospital use) and selecting any of them by cursor for immediate pronounciation,
- setting speech options during operation for changing speed, intonation, pitch, loudness etc.,
- reading prewritten text files,
- making conversation through telephone using the acoustic adapter,
- portable, approx. 60 min operating time powered by built-in battery,
- can be used as a LAP-TOP PC.

The basic research for MULTIVOX speech synthesis was made by Phonetics Laboratory in the Institute Liguistics of the Hungarian Academy of Sciences while applied research and development were carried out by Speech Research Laboratory Department of Telecommunications and Telematics, Techical University of Budapest. The manufacturer is NIKOL Electronics.

SPEECH RECOGNITION: VERBIDENT SYSTEMS

In the mid-eighties, an isolated word, limited vocabulary (appr. 200), speaker dependent system was developed (VERBIDENT-SD1). The improved second version is named VERBIDENT-SD2. It is implemented on an IBM-PC compatible computer equipped with a TMS32010 DSP plug-in board (developed at DTT, TUB). The system can work in conjunction with the keyboard. A flexible voice-command to keyboard-message emulator software facilitates application development. For increased speed an original method called geometric search was introduced. The system has a speaker independent version (VERBIDENT-SI) utilizing time-warped averaging a new scientific method.

A Hidden Markov model based connected word low vocabulary (approx. 30), speaker independent system is under investigation.

Recognition and text-to-speech systems have been integrated to form a two-way, speech based man-machine-interface.

SPEAKER IDENTIFICATION: SPI SYSTEM

In a speaker identification environment, the identification system stores a given speaker population (each represented by one or more portions of recorded speech), resulting in the unknown speaker's recorded speech. The task is to decide wether the unknown speaker is among the reference populations or not, and if so, which of the reference speakers is identical with the unknown one. SPI is an interactive speaker identification system with a software architecture which can handle a great number of recorded speech portion, and several affixed parameters. SPI integrates (under Microsoft Windows) graphical editor, with real-time speech record and listen facilities. Using this graphical editor user can extract different features both in the time and frequency domain. In addition SPI can automatically determine another kind of features relevant to a certain recorded speech. SPI's record manager handles speech records and features in an intelligent way so users have the possibility to quickly review and examine all features at any time. FInally, SPI compares the feature-vector of the unknown speaker with the feature-vector of all known speakers. The output is the identification of the speaker (if this is in the heap of the given population) using a weighted metric distance calculation. Weigthing can be easily changed by the user. It should be emphasized that SPI is an open architecture system to handle a great number of different phonetic (and other e.g. personnel) features, to form a feature-vector which is used for the decision making. System architecture is shown in Fig. 1.

The system components are grouped into 4 software modules which have individually icons in the Microsoft Windows framework as follows:

SPI Preprocess,

SPI Identification,
SPI Automatic,
SPI Maintenance.

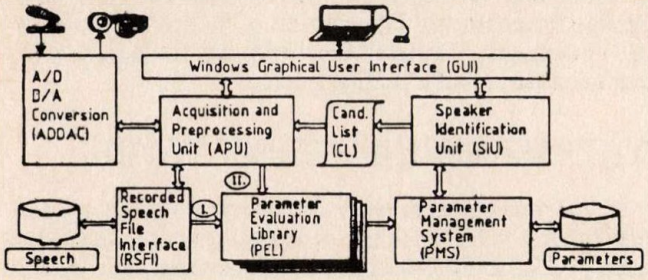


Fig. 1. Architecture of the speaker identification system

System requirements:

- IBM-PC/AT386 with i387 coprocessor, or 486,
- at least 4MB memory,
- at least 1 floppy drive
- min. 6MB hard disk capacity (depending on the amount of processed speech)
- EGA or better resolution and controller,
- mouse,
- analog interface board (optional: SPI can be used to process pre-recorded speech and parameters),
- Microsoft Windows 3.x

Supported analog interface boards:

- ARIEL DSP-16 (ARIEL Corp.),
- EBF rev.2. (TUB DTT),
- IBM analog interface adapter (IBM Corp.)
- UAM 5xx (Unielektroterv Gm.)

SPEECH QUALITY TEST SYSTEM: QUALIPHON

A speech quality test system called Qualiphon developed at DTT, TUB is capable of both simulating and testing low bit-rate speech coders and various digital and analog communication channels.

Speech communication services represent a wide variety and steadily growing application field for digital communication systems. The application include public commercial telephone networks, mobile communications, satellite communications, private communication lines, switched networks, cellular telephones, voice storage services etc. From an economical point of view, a communication channel should transmit as much information as possible, therefore efficient speech digitization and compression methods are needed. These are particularly important in the case of radio communication systems and voice storage applications where the bandwidth and information capacity are severely limited, necessitating low bit-rate speech coding methods. All speech encoding methods, however, have an undesirable side-effect, namely the degradation of speech quality. The fidelity of speech and the reduction of bit-rate are contradictory. The efficiency of speech digitization and compression can be measured directly by the resulting transmission bit-rate, but the fidelity of speech cannot be interpreted easily because of its subjective nature. Qualiphon is an integrated subjective speech quality assessment system which automates the process of subjective listening tests using an MNRU (Modulated Noise Reference Unit according to the CCITT Recommendation P.81.). The system has been developed for evaluating the subjective quality of low bit-rate speech codecs but it can be extended to include components for testing communication channels. Qualiphon has a speech sample library to be used for recording reference sentences. The samples to be evaluated can be recorded from the output of a hardware codec implementation, or can be the result of software simulation.

The system uses a Signal Processor Board (developed at DTT, TUB) plugged into an IBM PC/AT (286, 386 or 486) computer for implementing a real-time MNRU. The Qualiphon system is based on a DSPLab integrated digital signal processing environment.

SPEECH DETECTION: BD2 SYSTEM

This instrument was designed to detect the presence of human speech in noisy background with high sensitivity. Connecting a mixed speech and noise signal to the input of the speech detector a bistable output signalling is provided, its status depending on whether the segments of the input contain speech or not.

The loss of speech does not exceed 0.5% at a signal-to-noise ratio as low as 0...6 dB for average environmental noise. For special kind of noises (such as periodic signals) the performance is even better: more than 99.9% of the speech is detected safely.

GÁBOR MAGYAR
DTT/TUB

ON THE DESIGN AND REALIZATION OF WAVE DIGITAL FILTERS SATISFYING ARBITRARY AMPLITUDE SPECIFICATIONS

M. YASEEN* and T. HENK

DEPARTMENT OF TELECOMMUNICATIONS AND TELEMATICS
TECHNICAL UNIVERSITY OF BUDAPEST
H-1111 BUDAPEST, STOCZEK U. 2. HUNGARY

Design conceptions with corresponding structures are developed for wave digital filters satisfying arbitrary amplitude specifications in the passband and the stopband and possibly also in the transition band. Different forms of specifications e.g., monotonous and non-monotonous characteristics can be considered. The design conceptions rely on different wave digital filter structures, the first in one filter only, the second is a filter cascaded with an amplitude equalizer, the third consists of two cascaded filters. In the cascaded cases the first filter is constructed to satisfy the biggest part of the selectivity and can be restricted to be lattice structure or even bireciprocal one if the sampling frequency is considered as a free parameter while the second filter or equalizer ensures the final shape of the characteristic. The approximation problem is solved using interpolation methods combined with the Remez-exchange algorithm. Each design conception is optimized from the approximation and realization point of view. Illustrative examples are given to show the efficiency of each conception.

1. INTRODUCTION

Wave digital filters (WDF's) exhibit excellent properties, including coefficient sensitivity, roundoff noise, overflow level, stability and other nonlinear effects under finite-arithmetic conditions [1,2].

Robustness of wave digital filters is assured first of all by their passive reference filters [3] and specially the losslessness of the filter two-port [1. Sec. II. C]. The losslessness within passivity is automatically ensured if the approximated transmission is unity in the passband and is zero in the stopband. Otherwise there is no such direct relationship between losslessness and low coefficient sensitivity.

There are many applications, e.g., data transmission, sampling, etc. where the amplitude characteristics $A(\omega)$ is specified at all frequencies i.e., also in the transition band and the specified transmission is a given function in the passband.

Some examples for arbitrary passband and/or transition band specifications of a corresponding loss $\alpha(\omega) = -20\log(A(\omega))$ are shown in Fig. 1, which all fit well into the scope of the paper.

The purpose of this paper is to develop design conceptions by systematically considering different wave digital filter structures appropriate to this problem. Approximation methods for satisfying arbitrary amplitude specifications with these structures are presented. Each design conception is optimized from the design and realization point of view within its limitations. Typical examples are given to illustrate the efficiency of each conception.

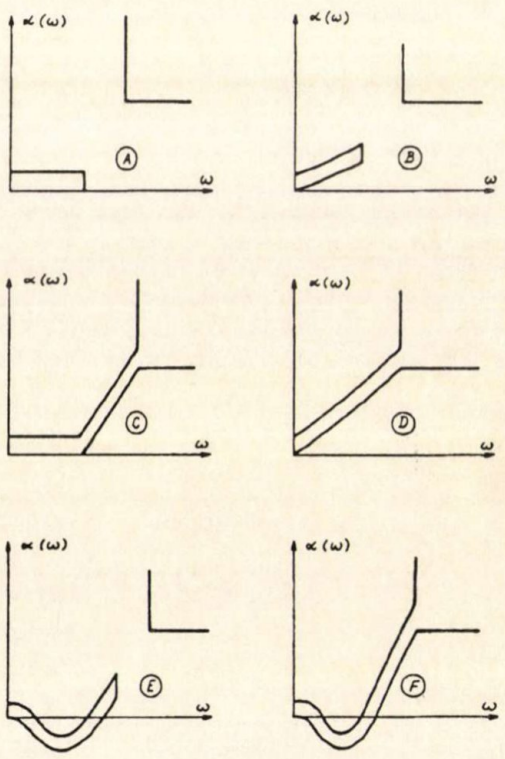


Figure 1. a: Prototype amplitude specification, b-f: arbitrary amplitude specification

2. BASIC WAVE DIGITAL FILTER STRUCTURES

To develop different design conceptions for solving the problem, some basic wave digital structures are recalled first. In this paper, structures with only real coefficients are considered. The characterization of these structures is given in the ψ reference frequency domain [1,2] and the classical network theory [4] is invoked (Table 1). These structures will be termed by the short name in the parentheses in the first column of Table 1. The scattering matrix [4] of lowpass, reciprocal structures are determined by the canonical polynomials $f(\psi)$, $g(\psi)$, and $h(\psi)$, where $g(\psi)$ is strictly Hurwitzian with degree n. The notation gg* refers to $g(\psi)g(-\psi) = f(\psi)f(-\psi)+h(\psi)h(-\psi)$, and

* On leave from Electrical and Electronic Engineering Dept. University of Assiut, Assiut, EGYPT

N denotes the number of free coefficients in $S_{21}(\psi)$. The abbreviation compl. stands for $S_{21}(\psi)S_{21}(-\psi) + S_{21}(1/\psi)S_{21}(-1/\psi) = 1$, the complementary property between the passband and the stopband in the bireciprocal case.

Table 1. Properties of lowpass reciprocal reference filters

Characterization	$f(\psi)$	$h(\psi)$	$g(\psi)$	$S_{21}(\psi)$	n	N
Reactant ladder (ladder)	even, with imaginary roots only	–	gg^*	–	–	$[3n+1]/2$ or $[3n]/2$
Reactant lattice	even	odd	gg^*	$\lvert S_{21}(0)\rvert = 1$	odd	n
Bireciprocal reactant lattice (bireciprocal)	even	odd	gg^* $g(\psi) = \psi^n g(1/\psi)$	$\lvert S_{21}(0)\rvert = 1$ compl.	odd	$[n-1]/2$
lossy lattice with dual canonic impedances (amplitude equalizer)	–	0	–	–	–	$2n+1$

3. DESIGN CONCEPTIONS

By considering systematically the basis wave digital structures, the developed design conceptions are summarized in Table 2. In the cascaded cases the first filter with higher order and attractive structure ensures the biggest part of the selectivity and the cascaded equalizer [5] or second filter with low order ensures the final shape of the amplitude characteristic in the passband and in the transition band. If the specification is given in form of lower and upper amplitude limit functions, the conditions on $A(\omega)$ are understood such that whether the specification can be overfullfilled in some frequency regions in order to satisfy the conditions at all frequencies. Bireciprocal filters can be employed if the sampling frequency F is considered as a free parameter for the filter design.

Table 2. Design Conceptions

Characterization	Conception	Conditions on the specification $A(\omega)$
One reactant filter	ladder	–
	lattice	$A(\omega) \leq A(0)$
	bireciprocal	$A(\omega) \leq A(0)$, free F, $A^2(\omega) + A^2(\pi.F - \omega) = A^2(0)$
One reactant filter cascaded with amplitude equalizer	lattice + equalizer	–
	birec. + eqalizer	free F
Two cascaded reactant filter	lattice + ladder	–
	birec. + ladder	free F
	lattice + lattice	$A(\omega) \leq A(0)$
	birec. + lattice	$A(\omega) \leq A(0)$, free F

4. APPROXIMATION AND REALIZATION

The approximation problem is solved for each design conception by applying interpolation methods combined with the Remez- exchange algorithm [6], while the properties of Table 1 are taken into consideration. Although delay specification is not considered now, the lattice and bireciprocal filters may become of non-minimum phase [7]. Each design conception is optimized from approximation and realization point of view within its limitations. If the specification is given in the digital frequency domain i.e., p or $(j\omega)$ domain, it is transformed to the reference frequency domain i.e., ψ or $(j\phi)$ domain through the relations of the bilinear transformation [1]:

$$\psi = \frac{z-1}{z+1} \ , \quad z = e^{PT} \ ,$$

$$\phi = \tan(\omega T/2) \ , \quad T = \text{sampling period.}$$

Bad) which are assigned by listeners. The measured quality is equal to the average value of scores received from all listeners.

In spite of being a good subjective measure of speech quality, MOS cannot distinguish fine differences between speech samples assessed by E-grade votes in case of high quality speech. To improve resolution at high quality level, the *Paired-Comparison Method* [4] has been developed. According to this method, two signals are presented to listeners who are asked to choose the better one. It is a forced comparison, "Equal" quality answer is not allowed. The percentage of the signal chosen as the preferred one is the preference score.

The *Equivalent Noise Paired Comparison Test* [4] is based upon a reference signal with varying signal-to-noise (SN) ratio. In this test, the quality is defined as the SN ratio of reference signal corresponding to 50% preference level. CCITT recommends a reference signal generator device called *Modulated Noise Reference Unit* (MNRU) for such tests in Recommendations P. 81 [5]. MNRU contains a white noise source modulated by the speech signal and the generated multiplicative noise is added to the speech signal to produce the reference signal (see Fig. 1). Such a reference signal has a speech component and a speech-amplitude correlated noise component with flat frequency spectrum. The signal-to-noise ratio (denoted by Q [dB]) can be set in the MNRU and is constant over the full dynamic range so its subjective effect is very similar to that of the distortion of logarithmic quantizers (standard PCM systems). Modifying MNRU with a noise-shaping filter results in a SN ratio which is almost independent from frequency.

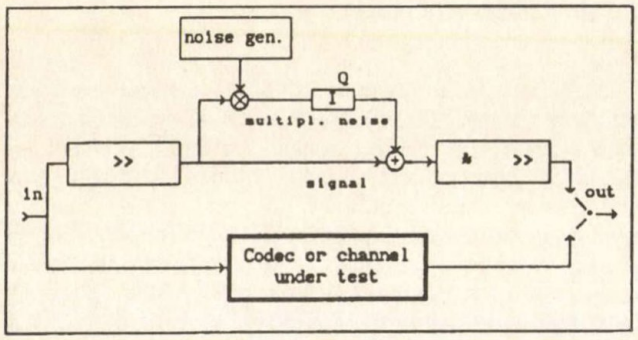


Fig. 1. Equivalent Noise Paired Comparison Test

Although it is sometimes difficult to compare signals with different types of impairments, the great advantage of the Paired Comparison Tests is that they provide highly accurate assessment on an absolute scale of speech quality even with as few as about 20 listeners. Therefore they are ideal for quick tests. The computer assisted speech quality assessment system Qualiphon developed at DTT, TUB and introduced in Sec. 4 is also based on the Equivalent Noise Paired Comparison Test. In contrary to this, quality tests based on MOS require hundreds or thousands of listeners to provide reliable results and the scale is relative, depending on time, country etc. like grades in schools. Nevertheless, thorough international investigations usually include MOS tests because it can take into account various kinds of deteriorations on a single scale.

2.2 Analytic methods

Analytic methods attempt to obtain different quality attributes of perceived speech by exploiting the phenomenon that listeners usually agree on the degree of speech impairment, but vary in their preference of that degradation. Therefore analytic methods generate a multidimensional characterization of the speech quality. Some methods have been developed which produce this kind of parametric description of speech such as *Paired Acceptability Rating Method (PARM), Quality Acceptance Rating Test (QUART) and Diagnostic Acceptability Measure (DAM)* [6].

For instance, in DAM a parametric scale is presented which is divided into three categories: signal category with ten rating scores, background category with seven rating scores and overall quality category with three rating scores. With the use of factor analysis, these twenty scores have been reduced to thirteen nearly independent perceptual quality scores. The signal category consists of the following parameters: fluttering, thinning, rasping, smothering, whining and irregularity. The background category consists of the following parameters: hissing, buzzing, bubbling and thumping. The general perceptual qualities are intelligibility, pleasantness and acceptability. From these parameters, overall attributes can be calculated such as total signal quality, total background quality and the most important factor which is based on all parameters, the composite acceptability.

Although the analytic methods provide a fairly good description of speech quality, such investigations are difficult and time consuming. No wonder that much effort has been paid to find objective speech quality measures which can efficiently predict subjective quality.

3. OBJECTIVE SPEECH QUALITY ASSESSMENT

However good assessment of speech quality can be provided by subjective tests, these have several disadvantages: they are expensive, slow, difficult to handle, non-repeatable due to the fact that human listeners' decisions depend on the test conditions and on their personal disposition. Especially the time-consuming nature of subjective measures excludes their use in the design and optimization of speech coding and communication systems.

Computable objective measures of speech quality based on measured physical parameters are much more desirable. They are cheap, simple, repeatable and fast in comparison with subjective measures, but they can be applied only if they predict subjective speech quality sufficiently well. So the task is to find an objective measure which can be efficiently computed from the original and distorted speech data set, and which highly correlates with subjective tests.

To solve this task is not easy because the human speech perception process is very complex and poorly understood. It involves also the grammar and other diverse factors such as the speakers' attitude and emotional state. People use a lot of redundant information in speech so as a result, certain slight distortion effects could cause complete intelligibility loss while other more extensive distortion products may be almost unperceivable. Quality assessment requires objective measures in order to take into consideration semantic, prosodic, syntactic, phonetic, etc. information. Of course, no objective measures

SPEECH QUALITY ASSESSMENT FOR LOW BIT-RATE CODING

S. MOLNÁR, P. TATAI and Z. JÁNOSY

DEPARTMENT OF TELECOMMUNICATIONS AND TELEMATICS
TECHNICAL UNIVERSITY OF BUDAPEST
H–1111 BUDAPEST, STOCZEK U. 2.

The paper reviews subjective and objective speech quality assessment methods. A particularly efficient method, the Paired Comparison Test based on a Modulated Noise Reference Unit (MNRU, CCITT Recommendation P.81) is described in detail. A speech quality test system called Qualiphon developed at DTT, TUB is also introduced. It is capable of both simulating and testing low bit-rate speech coders and various digital as well as analog communication channels. As an example of objective quality assessments using the built-in real time MNRU of the Qualiphon system, the procedure and some experimental results obtained with 4800 bit/s and a 3200 bit/s CELP coder are presented.

1. INTRODUCTION

Speech communication services represent a wide variety and steadily growing applications field for digital communication systems. The applications include public commercial telephone networks, mobile communications, satellite communications, private communication lines, switched networks, cellular telephones, voice storage services etc. From an economical point of view, a communication channel should transmit as much information as possible, therefore efficient speech digitization and compression methods are needed. These are particularly important in the case of radio communication systems and voice storage applications where the bandwidth and information capacity are severely limited, necessitating low bit-rate speech coding methods. In the following, low bit-rate means rates below the standard value of 64 Kbit/s.

All speech coding methods, however, have an undesirable side-effect, namely the degradation of speech quality. The fidelity of speech and the reduction of bit-rate are contradictory. The efficiency of speech digitization and compression can be measured directly by the resulting transmission bit-rate but the fidelity of speech cannot be interpreted easily because of its subjective nature. Some interpretations will be introduced in this paper. In order to evaluate speech coding systems, speech quality assessment methods are needed [1]. These are very important not only for optimizing coding algorithms but also for designing effective communication systems. There are two categories of such methods: subjective and objective speech quality assessments. The subjective methods are based on standardized procedures which use humans to judge the quality of speech. In contrast, the objective methods eliminate human judgements from the assessment procedure and provide computable results based on measurable physical quantities. The main problem of finding a good objective speech quality assessment method, however, is that its results should highly correlate with users' opinion, so once again one has to resort to subjective test in order to "calibrate" objective measures.

The main goal of this paper is to give a brief summary of subjective and objective speech quality assessment methods and to introduce the Qualiphon system which is an efficient speech quality assessing frame program based on the DSPLab signal processing environment developed at DTT, TUB.

2. SUBJECTIVE SPEECH QUALITY ASSESSMENT

Speech quality depends primarily on human perception so subjective quality assessment methods imply humans as referees. There are two categories of subjective measures: *utilitarian* and *analytic*. Utilitarian methods measure speech quality on a unidimensional scale so that results can be summarized by a single number capable of comparing communication systems directly. Analytic methods generate their results on a multidimensional scale reflecting various speech quality components.

2.1. Utilitarian methods

2.1.1. Intelligibility tests

There is a wide range of utilitarian methods used extensively mainly for very low bit-rate or synthetic voice. One category focused on speech intelligibility called *Intelligibility Tests* consists of articulation tests, rhyme tests and speech interference tests [2]. *Articulation Tests* [3] give their results as a percentage of correctly received sounds, words and sentences. The rate of correctly heard monosyllables is known as syllabic articulation. Sound articulation is defined as a percentage of correctly received phonemes. With modifying articulation tests the *Equivalent Loss Method* [3] is obtained. The score of this method is the Articulation Equivalent Loss (AEN) which is defined as the difference in attenuation values at 80% sound articulation between reference and test system.

2.1.2. Quality tests

Another category of utilitarian methods comprises *Quality Tests*. The intelligibility tests are unable to measure the speech quality when speech is highly intelligible, and this is exactly what we expect from most speech services. So methods are needed which can measure other attributes such as pleasantness or naturalness. For these purposes new methods have been developed. The most widely used method based on opinion rating is the *Mean Opinion Score* (MOS) [3]. In this test, five grades of speech quality are usually distinguished (Excellent, Good, Fair, Poor and

The approximation procedure is summarized for each design conception as follows.

4.1. One reactant filter

According to this conception, the given amplitude specification is approximated by only one reactant filter. In case of $A(\omega) \leq A(0)$, the designed filter can be restricted to be odd-degree lattice or even bireciprocal one if the sampling frequency is considered as a free parameter. The design itself is carried out by constructing the characteristic function $\Psi(\psi)$ [1].

$$\Psi(\psi) = \frac{h(\psi)}{f(\psi)} = \frac{S_{11}(\psi)}{S_{21}(\psi)} \quad ,$$

where $S_{11}(\psi)$ is the reflectance and $S_{21}(\psi)$ is the transmittance.

In case of lattice structure, $h(\psi)$ is an odd polynomial with degree n, which is the filter degree and it is an odd number:

$$h(\psi) = \psi \sum_{i=0}^{\frac{n-1}{2}} h_i \psi^{2i} \quad ,$$

and its roots are not restricted to be on the $(j\phi)$ axis. On the other hand, $f(\psi)$ is an even polynomial of degree n-1 and can be formulated as:

$$f(\psi) = \sum_{i=0}^{\frac{n-1}{2}} f_i \psi^{2i} \quad ,$$

and it is restricted to possess roots on the $(j\phi)$ axis only, i.e.,

$$f(\psi) = \prod_{i=1}^{\frac{n-1}{2}} (\psi^2 + B_i^2) \quad ,$$

where B_i are the corresponding transmission zeros. The specification for $|\psi(j\phi)|$ is obtained from that of $S_{21}(\psi)$:

$$|\Psi(j\phi)|^2 = 1/|S_{21}(j\phi)|^2 - 1.$$

The design procedure can now be summarized as:

1) Calculate the specification of the characteristic function from the specification of the transmittance and select a filter degree.
2) Select proper values for the transmission zeros B_i.
3) At a selected set of frequencies, interpolate $h(\psi)$ for h_i in the passband and transition band if it is also specified.
4) Apply Remez-exchange algorithm to move the interpolation frequencies for more optimal h_i.
5) With the obtained h_i, apply Remez-exchange algorithm to move frequencies B_i for more optimal $f(\psi)$.
6) Go back to step 3 to re-satisfy the passband and transition band. Continue till the specification is satisfied in the passband, stopband and also in the transition band if specified. If the degree of the filter is not sufficient to satisfy the stopband, select new degree for the filter and repeat from step 2.

In case of bireciprocal structure design, the function $\Psi(\psi)$ is decomposed as:

$$\Psi(\psi) = \Psi_1(\psi)\Psi_2(\psi),$$

where $\Psi_1(\psi)$ is restricted to possess roots on the $(j\phi)$ axis only, while $\Psi_2(\psi)$ is not restricted to possess roots on the $(j\phi)$ axis:

$$\Psi_1(\psi) = \frac{\psi \prod_{i=1}^{k/2} (A_i^2\psi^2 + 1)}{\prod_{i=1}^{k/2} (\psi^2 + A_i^2)},$$

$$\Psi_2(\psi) = \frac{\sum_{i=0}^{m/2} r_i \psi^{2i}}{\psi^m \sum_{i=0}^{m/2} r_i \psi^{-2i}},$$

where, $k+m = n-1$ and n is the filter degree. This form of $\Psi_1(\psi)$ and $\Psi_2(\psi)$ guarantees the bireciprocal property [7]:

$$\Psi(1/\psi) = 1/\Psi(\psi).$$

The approximation procedure can be summarized as:

1) Calculate the specification for the characteristic function from the specification of the transmittance and select a filter degree.
2) Select proper values for transmission zeros A_1.
3) At a selected set of frequencies, interpolate $\Psi_2(\psi)$ for r_i in the passband and transition band if specified.
4) Apply Remez-exchange algorithm to move the interpolation frequencies for more optimal r_i.
5) With the obtained r_i, apply Remez-exchange algorithm to move frequencies A_i for more optimal $\Psi_1(\psi)$.
6) Go back to step 3 to re-satisfy the passband and transition band. Continue till the specification is satisfied in the passband, stopband and also in the transition band if specified. If the filter degree is not sufficient to satisfy the stopband select new degree and repeat from step 2.

In case of $A(\omega) > A(0)$ in some frequency region, the designed filter must be of ladder structure. In this case the transmittance $S_{21}(\psi)$ can be expressed as:

$$S_{21}(\psi) = \frac{f(\psi)}{g(\psi)} = \frac{\prod_{i=1}^{m/2} (\psi^2 + C_i^2)}{\sum_{i=0}^{n} g_i \psi^i} \quad ,$$

where $f(\psi)$ is an even polynomial of degree m with $(j\phi)$ axis roots only. On the other hand $g(\psi)$ is a strictly Hurwitzian polynomial of degree n which is the filter degree and can be even or odd number. The approximation is achieved through the construction of the squared magnitude transmittance:

$$S_{21}(\psi)S_{21}(-\psi) = \frac{\prod_{i=1}^{m/2} (\psi^2 + C_i^2)^2}{\sum_{i=0}^{n} d_i \psi^{2i}} = \frac{N(\psi)}{D(\psi)}, \psi = j\phi.$$

The approximation procedure can be summarized as:

1) Assess the degrees for $f(\psi)$ and $g(\psi)$.
2) Specify suitable values for the transmission zeros C_i and consequently $N(\psi)$ and $f(\psi)$ are determined.
3) At a selected set of frequencies, interpolate $D(\psi)$ for the coefficients d_i.

4) Apply Remez-exchange algorithm to change the interpolation frequencies for more optimal d_i.
5) With the obtained d_i, apply Remez-exchange algorithm for more optimal C_i.
6) Go back to step 3 to re-satisfy the passband and transition band. Continue till the specification is satisfied in the passband, stopband and the transition band if specified. If the selected degree is not sufficient to satisfy the stopband, select new degrees and repeat from step 2.
7) Get the Hurwitzian factorization of $D(\psi)$ to construct $g(\psi)$.

To realize a lattice filter from its characteristic function, the expression:

$$h(\psi) + f(\psi) = g_1(\psi)g_2(-\psi)$$

is considered [1,2,4], where $g_1(\psi)$ and $g_2(\psi)$ are strictly Hurwitzian polynomials, one of them with odd degree and the other with even degree, and they are related to the polynomial $g(\psi)$ by

$$g(\psi) = g_1(\psi)g_2(\psi).$$

So, $g_1(\psi)$ and $g_2(\psi)$ are calculated by taking Hurwitz and anti-Hurwitz factors of $h(\psi) + f(\psi)$, respectively. Then, the two branch all-pass functions $S_1(\psi)$ and $S_2(\psi)$ are defined:

$$S_1(\psi) = -g_1(-\psi)/g_1(\psi), S_2(\psi) = g_2(-\psi)/g_2(\psi).$$

Consequently, the transmittance $S_{21}(\psi)$ can be obtained:

$$S_{21}(\psi) = [S_2(\psi) - S_1(\psi)]/2 \quad .$$

The two all-pass functions $S_1(\psi)$ and $S_2(\psi)$ are realized through their factorization into first and second order cascaded sections:

$$S_k(\psi) = \prod_j \frac{-\psi + X_{j,k}}{\psi + X_{j,k}} \prod_i \frac{\psi^2 - Y_{i,k}\psi + Z_{i,k}}{\psi^2 + Y_{i,k}\psi + Z_{i,k}}, k = 1, 2$$

and consequently, the corresponding multiplier coefficients for the WDF structure of Ref. [2] are:

$$\gamma_j = \frac{1 - X_{j,k}}{1 + X_{j,k}}, \gamma_{2i-1} = \frac{Y_{i,k} - Z_{i,k} - 1}{Y_{i,k} + Z_{i,k} + 1}, \gamma_{2i} = \frac{1 - Z_{i,k}}{1 + Z_{i,k}}.$$

In the case of bireciprocal filter design, $g_1(\psi)$ and $g_2(\psi)$ occur in quadratic root arrangement forms, so, to realize them using cascaded first and second order all-pass sections, we have to choose two dummy variables [8] $\psi^{\sim}$ and $z^{\sim}$, which are related to the original variables ψ and z by:

$$\psi^{\sim} = [\psi + 1/\psi]/2, \quad 1/z^{\sim} = -z^{-2}, \quad \psi^{\sim} = [z^{\sim} - 1]/[z^{\sim} + 1].$$

In case of ladder filters realization, we use Fujisawa's theorems [9,10] to realize it as either lowpass midseries or lowpass midshunt lossless two-port terminated with a resistance.

4.2. Reactant filter cascaded with an amplitude equalizer

According to this conception, the given specification is approximated such that an attractive structure filter with a prototype [10] amplitude specification (Fig.1a) is designed first to satisfy the biggest part of the selectivity. This filter can be restricted to possess odd degree lattice structure or even a bireciprocal one if the sampling rate is considered as a free parameter. Explicit formulas [2] can be used to design this filter. After this, the characteristic of the designed filter is subtracted from the original specification to get new upper and lower amplitude limit functions for the specification of the required equalizer.

The equalizer is approximated through the construction of its squared magnitude transmittance:

$$S_{21}(\psi)S_{21}(-\psi) = \frac{f(\psi)f(-\psi)}{g(\psi)g(-\psi)} =$$

$$= \frac{\sum_{i=0}^{m} n_i\psi^{2i}}{\sum_{i=0}^{m} d_i\psi^{2i}} = \frac{N(\psi)}{D(\psi)} \quad , \psi = j\phi \quad ,$$

where $g(\psi)$ is a strictly Hurwitzian polynomial and there is no restriction for $f(\psi)$ and both of them are of degree m.

The approximation procedure is summarized as:
1) Assess the degrees for the filter and the equalizer.
2) Design first a prototype lattice (or bireciprocal) filter with higher order to satisfy the biggest part of the selectivity.
3) Get the specification for a low order amplitude equalizer.
4) At a selected set of frequencies, interpolate the squared magnitude transmittance of the equalizer for n_i, d_i.
5) Apply Remez-exchange algorithm to move the interpolation frequencies for more optimal n_i, d_i.
6) Design more optimal lattice filter and repeat from step 3 till the resulting response satisfies the specification.
7) If the stopband is not or is over satisfied, select new degrees and repeat from step 2.
8) The transmittance $S_{21}(\psi)$ of the required equalizer is constructed through any factorization of $N(\psi)$ and the Hurwitz factorization of $D(\psi)$.

To realize the equalizer, its transmittance $S_{21}(\psi)$ is equal to the reflectance $S_{11}^c(\psi)$ of a lossy one- port reference circuit with input impedance $Z_{in}(\psi)$ when a circulator is applied [1,5]:

$$S_{21}(\psi) = S_{11}^c(\psi) = [Z_{in}(\psi) - 1]/[Z_{in}(\psi) + 1].$$

where unity reference resistance is assumed.

The realization is carried out by first and second order cascaded sections. The general reference circuit structure of a second order amplitude equalizer is shown by Fig. 2. The reflectance of this circuit can be expressed as:

$$S_{11}^c(\psi) = k_d \frac{\psi^2 + a\psi + b}{\psi^2 + c\psi + d} \quad ,$$

where k_d, a, b, c and d are obtained from the equalizer approximation. If the admittance of this equalizer circuit is written as:

$$Y(\psi) = \frac{\alpha_1\psi^2 + \beta_1\psi + \varsigma_1}{\alpha_2\psi^2 + \beta_2\psi + 1} \quad ,$$

then, the element values can be determined from the following set of equations:

$$LC/R_1 + LC/R_3 = \alpha_1$$
$$C + L/R_1R_2 + L/R_2R_3 + L/R_1R_3 = \beta_1$$
$$1/R_2 + 1/R_3 = \varsigma_1$$
$$LC = \alpha_2$$
$$L/R_1 + L/R_2 = \beta_2$$

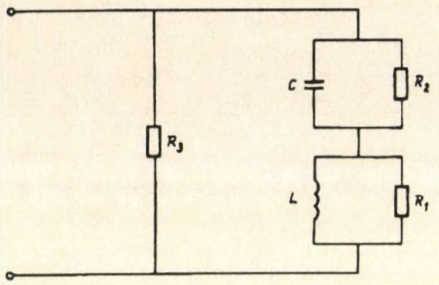


Figure 2. Reference circuit of the second order amplitude equalizer

If positive values are restricted for the elements, k_d is replaced by $k_r, k_d > k_r$. This exhibits a few dB insertion loss due to the equalizer which is irrelevant from the practical point of view. An interesting choice for k_r yields more simplification in the equalizer circuit, for example R_2 becomes infinite if k_r is set to:

$$k_{r\infty} = \frac{c^2b - cad - db + d^2}{b^2 - db + a^2d - cba}$$

Another optimization for the equalizer realization can be achieved if $A(\omega) \leq A(0)$. In this case the corresponding one-port can be simplified to the structure shown in Fig. 3. In such a case the one-port can be interpreted as a lossless ladder two-port network terminated by a resistance i.e., as a ladder filter. The design of this network can be achieved through the construction of its transmittance rather than its reflectance:

$$|S^c_{21}(j\phi)|^2 = 1 - |S^c_{11}(j\phi)|^2, \quad S^c_{11}(\psi) = S_{21}(\psi).$$

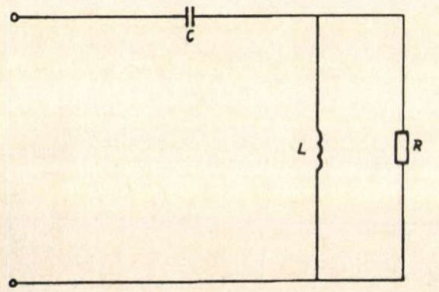


Fig. 3. Simplified reference circuit of second order amplitude equalizer

The transmittance $S^c_{21}(\psi)$ of the lossless two-port can be formulated now as:

$$S^c_{21}(\psi) = \frac{f^c(\psi)}{g(\psi)} = \frac{\psi^2}{u_0\psi^2 + u_1\psi + u_2} \quad .$$

The approximation is achieved through the construction of the squared magnitude transmittance of the lossless two-port:

$$S^c_{21}(\psi)S^c_{21}(-\psi) = \frac{\psi^4}{v_0\psi^4 + v_1\psi^2 + v_2} \quad , \quad \psi = j\phi \quad .$$

After getting the values of the coefficients v_i, $S^c_{21}(\psi)$ can be calculated, and the values R, L, C can be determined from the following set of equations:

$$[R+1]^2/4R = v_0$$
$$[R+1]/2LC - [L + RC]^2/4RL^2C^2 = v_1$$
$$R/4L^2C^2 = v_2$$

4.3. Two cascaded reactant filters

According to this conception the given specification is approximated such that the first filter with attractive structure and with a prototype amplitude specification is designed to satisfy the biggest part of the selectivity. This filter can be restricted to possess odd degree lattice structure or even bireciprocal one if the sampling frequency is considered as a free parameter. Explicit formulas can be used [2] to design this filter. The characteristic of the first filter is subtracted from the original amplitude specification to get the specification for the second filter which is approximated by using the interpolation method combined with Remez-exchange algorithm similarly to the conception of one filter. If $A(\omega) \leq A(0)$, the second filter can be also a lattice filter, otherwise, it has to be a ladder type. If the second filter has to be of ladder structure, it can be polynomial type filter, i.e., without finite transmission zeros.

The presented design conceptions are compared from the realization point of view through illustrative examples.

5. ILLUSTRATIVE EXAMPLES

A monotonous amplitude specification $(A(\omega) \leq A(0))$ like the characteristic of a data transmission demodulator filter will be approximated using the above explained design conceptions.

The obtained results are tabulated as follows:

* Table 3 compares the design conceptions.
* Table 4 gives the numerical results.

The column no. 6 in Table 3 refers to the loss reserve in the stopband. Since the stopband is not equiripple in most cases, the reserve is expressed by the minimal and maximal ripple. The same column refers also to some Figures which illustrate the loss characteristic together with the given specification for the solutions with minimum number of adaptors. The last column in Table 3 refers to Figs. 6–11 illustrating the designed WDF structures. The corresponding γ_i or $\gamma_{i,j}$ coefficients are given in Table 4. From the middle column of Table 4, the following features are observed:

1) The one 7-th order lattice solution has three transmission zeros and no reflection zeros on the $(j\phi)$ axis.
2) The one 9-th order bireciprocal solution has two transmission and reflection zeros on the $(j\phi)$ axis.
3) The first lattice (or bireciprocal) filter in the cascaded cases of degree n_1 has $[n_1 - 1]/2$ transmission and reflection zeros on the $(j\phi)$ axis.

4) The second filter in the two filters solution has one transmission zero and no reflection zeros on the $(j\phi)$ axis.

5) The amplitude equalizer has no transmission zeros on the $(j\phi)$ axis.

Table 3. Comparison of the design conceptions

Design conception	Degree	F KHz	No. of adaptors	No. of multipliers	Loss reserve in stopband, (loss response and specification)	WD realiz-ation
Lattice	7	64.1	7	7	2.2-11.5 dB	Fig. 6
Bireciprocal lattice	9	63.068	4	4	1.4-1.5 dB (Fig. 4)	Fig. 7
Lattice + L,C equilizer of Fig. 3	5 + 2=7	64.1	5 + 2=7	5 + 3=8	4.8-5.9 dB	Fig. 8
Bireciprocal + L,C equilizer of Fig. 3	7 + 2=9	80.0	3 + 2=5	3 + 3=6	6.8-22 dB	Fig. 9
Lattice + lattice	3 + 3=6	64.1	3 + 3=6	3 + 3=6	2.7-5.7 dB (Fig. 5)	Fig. 10
Bireciprocel + lattice	5 + 3=8	80.0	2 + 3=5	2 + 3=5	6.5-21.4 dB	Fig. 11

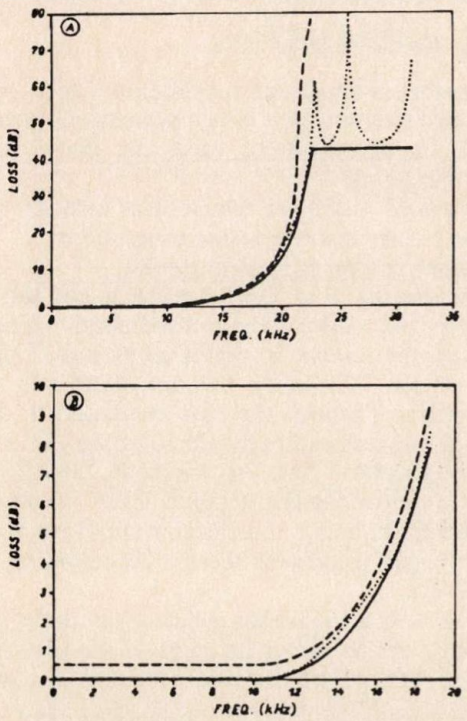


Fig. 4. Loss response of bireciprocal filter construction (with the specification)

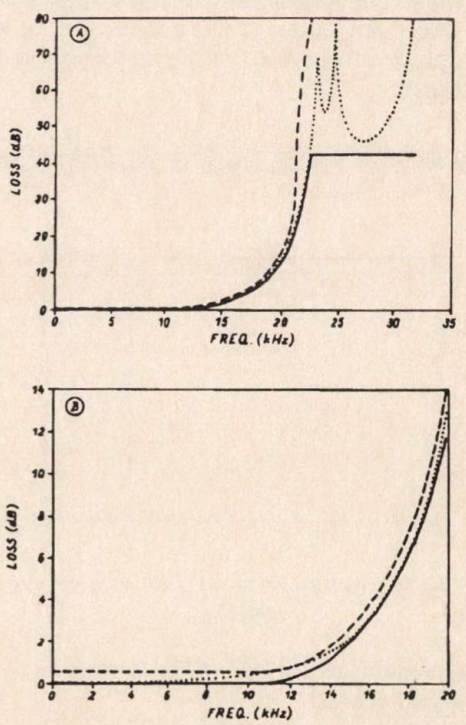


Fig. 5. Loss response of lattice + lattice filter construction (with the specification)

Table 4. Numerical results

Design conception	$f(\psi), g(\psi)$ or $g_1(\psi), g_2(\psi)$	γ_i or $\gamma_{i,j}$ i	j	Value
Lattice	$g_1(\psi) = (\psi + 1.777225)$	0		-0.27985669
	$(\psi^2 + 1.830517\psi + 3.36203)$	1		-0.39024214
	$g_2(\psi) = (\psi^2 + 1.208118\psi + 1.754639)$	2		-0.273952
	$(\psi^2 + 2.367382\psi + 2.703734)$	3		-0.40879996
		4		-0.54149788
		5		-0.2201163
		6		-0.4600044
Bireciprocal	$g_1(\psi) = (\psi + 1)(\psi^4 + 1.320034\psi^3 +$	1		-0.11089074
	$+2.80987135\psi^2 + 1.320034\psi + 1)$	2		0.417039784
	$g_2(\psi) = (\psi^4 + 2.259236\psi^3 +$	3		-0.291251246
	$+3.6455719\psi^2 + 2.259236\psi + 1)$	4		0.663246149
Lattice	$g_1(\psi) = (\psi + 0.9291056)$	0		0.0367498
	$(\psi^2 + 0.348789\psi + 1.911421)$	1		-0.3171611
	$g_2(\psi^2 + 0.348789\psi + 1.911421)$	2		-0.1536144
		3		-0.7860327
+		4		-0.3130502
	$f^c(\psi) = \psi^2$	1	1	0.22664185
	$g(\psi) = 1.9403241(\psi^2 + 0.856266\psi +$	1	2	1.2042978
L,C eqalizer	$+1.2471912)$	1	3	0.56980603
		2	2	0.6673958
		2	3	0.3326042
Bireciprocal	$g_1(\psi) = (\psi + 1)(\psi^2 + 0.7980506\psi + 1)$	1		-0.1284564
	$g_2(\psi) = (\psi^2 + 1.544665\psi + 1)$	3		-0.4295667
+	$(\psi^2 + 0.2337946\psi + 1)$	5		-0.7906749
	$f^c(\psi) = \psi^2$	1	1	0.321929
L,C equalizer	$g(\psi) = 1.1695289(\psi^2 + 0.843576\psi +$	1	2	0.6780707
	$+0.708861)$	2	1	0.5949425
		2	2	0.8191454
		2	3	0.585912
Lattice	$g_1(\psi) = \psi + 1.424332$	0		-0.1750306
	$g_2(\psi) = \psi^2 + 0.7201682\psi + 2.203382$	1		-0.6328998
+		2		-0.3756599
	$g_1(\psi) = \psi + 1.504977$	0		-0.201589
	$g_2(\psi) = \psi^2 + 1.234707\psi + 1.244052$	1		-0.2901451
lattice		2		-0.108755
Bireciprocal	$g_1(\psi) = (\psi + 1)(\psi^2 + 0.237982\psi + 1)$	1		-0.3123584
+	$g_2(\psi) = \psi^2 + 1.047948\psi + 1$	3		-0.7873244
	$g_1(\psi) = \psi + 0.8510658$	0		0.0804586
lattice	$g_2(\psi) = \psi^2 + 0.7510658\psi\psi + 0.5408208$	1		-0.3445872
		2		0.298009645

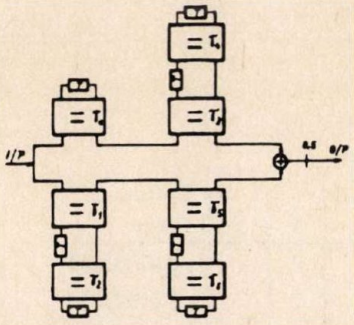


Fig. 6. WD realization of lattice filter construction

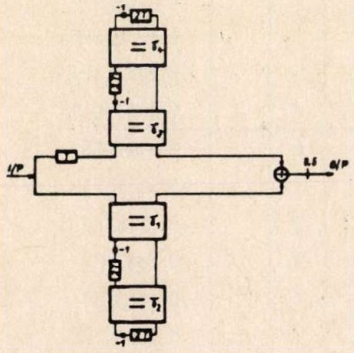


Fig. 7. WD realization of bireciprocal filter construction

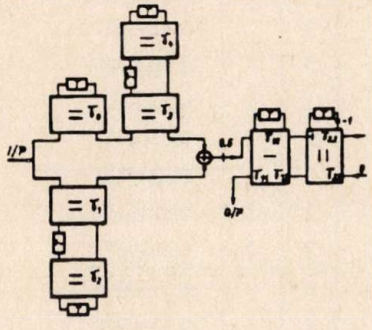


Fig. 8. WD realization of lattice + L,C equalizer filter construction

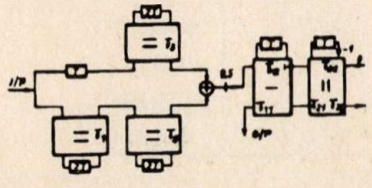


Fig. 9. WD realization of bireciprocal + L,C equalizer filter construction

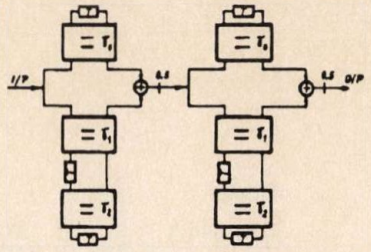


Fig. 10. WD realization of two lattices filter construction

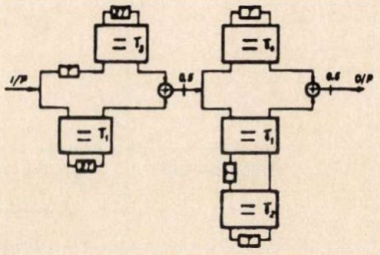


Fig. 11. WD realization of bireciprocal + lattice filter construction

6. CONCLUSIONS

Several design conceptions are presented to approximate arbitrary amplitude specifications. If the sampling frequency F is considered as a free parameter for the filter design, the whole filter or a part of it can be restricted to be bireciprocal structure yielding more save in the WD realization. In such a case the conception of one bireciprocal filter is the most optimal one from the WD realization point of view since the bireciprocal property reduces the number of adaptors in every part of the filter structure. If on the other hand no freedom is given for the sampling frequency F, the structure of two odd degree lattice filters gives the best solution in our example with a total even degree which is less by one than the odd-degree of the corresponding one lattice structure. This result is very useful in practice since a real coefficient lattice filter can not be designed with an even degree which may be sufficient to satisfy the specification. The lossy equalizer results in more attenuation in the stopband, but it also exhibits small insertion loss. This insertion loss disappears in case of monotonous specification by using L.C equalizers which give zero insertion loss but also give smaller attenuation in the stopband. These solutions require however higher number of adaptors.

Further simulation studies will be necessary to compare these design conceptions and structures from sensitivity, scaling, dynamic range and nonlinearity point of view. This will be included in a further contribution.

ACKNOWLEDGEMENTS

The authors would like to express their thanks to Ferenc Leeb, Julianna Földvári-Orosz, László Hinsenkamp and Lajos Fürjes for their valuable help. Thanks are also given to the computer staff of the Research Inst. for Telecommunications and of the Dept. of Telecommunications and Telematics, Tech. Univ. of Budapest for supplying the computer facilities.

REFERENCES

[1] A. Fettweis, "Wave Digital Filters: Theory and Practice", Proc. of IEEE, vol. 74. February 1986, pp. 270-327.

[2] L. Gazsi, "Explicit Formulas for Lattice Wave Digital Filters", IEEE Trans. Circuits and Systems, vol. CAS-32, January 1985, pp. 68-88.

[3] A. Fettweis, "Some General Properties of Signal-Flow Networks", Network and Signal Theory edited by J.K. Skwirzynski and J.O. Scanlan, Peter Peregrinus Ltd, London 1973, pp. 48-59.

[4] V. Belevitch, *Classical Network Theory*, Holden-day, 1968.

[5] U. Sauvagard, "A Ten-Channel Equalizer for Digital Audio-Applications", IEEE Trans. Circuits and Systems, vol. CAS-36, February 1989, pp. 276-280.

[6] J. Földvári-Orosz, T. Henk, E. Simonyi, "Simultaneous Amplitude and Phase Approximation for Lumped and Sampled Filters", International Journal of Circuit Theory and Applications, vol. 19, pp. 77-100, Jan-Feb. 1991.

[7] F. Leeb, T. Henk, "Simultaneous Approximation for Bireciprocal Lattice Wave Digital Filters", Proc. of the ECCTD'89, Brighton U.K., 5-8 Sept. 1989, pp. 472-476.

[8] W. Wegener, "Wave Digital Filters with Reduced Number of Multipliers and Adders", AEU, Band 33, 1979, Heft 6, pp. 239-243.

[9] H. Baher, *Synthesis of Electrical Networks*, John Wiley and Sons, 1984.

[10] J.D. Rhodes, *Theory of Electrical Filters*, Willey, 1976.

Mohamed Yaseen received the B.Sc. and M.Sc. degrees of Electrical Engineering (Communications) in 1978 and 1984, respectively from Assiut University, Egypt. His M.Sc. research was concerned with the digital signal processing, typically: time domain design of recursive digital filters. He worked as a demonstrator and then as assistant lecturer at the Electrical and Electronic Eng. Dept., Assiut University, Egypt. In 1988 he got a scholarship to study for the Ph.D. degree in Hungary. He was with the Research Institute for Telecommunications from 1988 to 1990. Now he is with the Dept. of Telecommunications and Telematics, Tech. Univ. of Budapest. His research activities are within the design, realization and simulation of digital filters.

Tamás Henk received the M.Sc. and Dr. Tech. Electronic Engineering degrees from the Tech. Univ. of Budapest, Hungary in 1973 and 1980, respectively, and the Ph.D. degree from the Hungarian Academy of Sciences in 1985. He was with the Research Institute for Telecommunications from 1973 to 1990 practising in development and management and is presently an associate professor at the Dept. of Telecommunications and Telematics, Tech. Univ. of Budapest. He was a postdoctoral research fellow at the Univ. College, Dublin, Ireland from 1977 to 1979. His research interests are within circuit theory, digital signal processing and telecommunications.

BOOK REVIEW

PROTOCOL SPECIFICATION AND TESTING by Katie TARNAY

Our world is moving towards what we might call the information society. In this information society everybody should have an up-to-date computing environment with an easy-to-achive communication background. The principal means of integrating the computer environment and the communication tools is computer communication. Computer communication is responsible for the data message exchange between the computerized source station and the destination station.

The book deals with the communication protocols. First the reference model and its layers are discussed, then the seven layers and their protocols are described. There are the specifications of BSC, DLC and X.25 protocols. The author defines the abstract model of testing.

In the second part the reader can find the formal description techniques: state-transition based models, graph models, algebras and formal languages, specification languages: SDL, ESTELLE, LOTOS.

The third part deals with applications. The protocols introduced in the first part are specified using the methods of the second part.

For applications the user can read the alternating-bit protocol specification with state-transition graph, Petri-net and with ESTELLE. There is the formal description of HDLC and the LAPB with SDL/GR. For the transport protocols there are case studies in SDL-graph, numerical Petri-nets, data flow graph and LOTOS specification.

With connection of conformance testers and test sequences we can read about the test sequence generation, w-method, Gönenc method, the automatic test sequence generation and about case studies for the testing the NATHAN application protocol, and about Conformance Test Centre in the National Bureau of Standards.

The reader will be familiarized in computer communication, protocol specification with formal descripton techniques and protocol testing.

The chapters contain examples and the most important international data communication standards, literature and abbreviations are summarized at the end of the book.

The book is suitable for self-study and for university course book.

The book was published concurrently by Akadémiai Kiadó, Budapest, Hungary and Plenum Press, New York, USA.

GY. CSOPAKI

Information for authors

JOURNAL ON COMMUNICATIONS is published monthly, alternately in English and Hungarian. In each issue a significant topic is covered by selected comprehensive papers.

Other contributions may be included in the following sections:

- INDIVIDUAL PAPERS for contributions outside the focus of the issue,
- PRODUCTS-SERVICES for papers on manufactured devices, equipments and software products,
- BUSINESS-RESEARCH-EDUCATION for contributions dealing with economic relations, research and development trends and engineering education,
- NEWS-EVENTS for reports on events related to electronics and communications,
- VIEWS-OPINIONS for comments expressed by readers of the journal.

Manuscripts should be submitted in two copies to the Editor in chief (see inside front cover). Papers should have a length of up to 30 double-spaced typewritten pages (counting each figure as one page). Each paper must include a 100–200 word abstract at the head of the manuscript. Papers should be accompanied by brief biographies and clear, glossy photographs of the authors.

Contributions for the PRODUCTS-SERVICES and BUSINESS-RESEARCH-EDUCATION sections should be limited to 16 double-spaced typewritten pages.

Original illustrations should be submitted along the manuscript. All line drawings should be prepared on a white background in black ink. Lettering on drawings should be large enough to be readily legible when the drawing is reduced to one- or two-column width. On figures capital lettering should be used. Photographs should be used sparingly. All photographs must be glossy prints. Figure captions should be typed on a separate sheet.

For contributions in the PRODUCTS-SERVICES section, a USD 110 page charge will be requested from the author's company or insitution.

EUROPA TELECOM'92

Budapest, 12 – 17 October, 1992

The most importantly timed regional event ever organized by the International Telecommunication Union (ITU). The Exhibition and Special Session of the World Telecommunication Forum including Policy, Technical and Economic Symposia that will shape the future of Europe.

Taking place in Budapest, the crossroads of the two "economic civilisations".

A unique opportunity to meet with the leaders of major companies, learn about critical legislation issues, discuss joint ventures and investments.

CALENDAR OF EUROPA TELECOM'92
Budapest, 12–17 October, 1992

Sunday 11 October	Monday 12 October	Tuesday 13 October	Wednesday 14 October	Thursday 15 October	Friday 16 October	Saturday 17 October
ITU PRESS DAY	OFFICIAL OPENING CEREMONY 10.00–11.30	EXHIBITION Opening hours 9.30–17.30	EXHIBITION Opening hours 9.30–17.30	EXHIBITION Opening hours 9.30–17.30	EXHIBITION Opening hours 9.30–17.30	EXHIBITION Opening hours 9.30–17.30
TELECOM'92 Official Receptioln	EXHIBITION Opening hours 12.00–17.30					
	POLICY SYMPOSIUM 13.45–17.30	POLICY SYMPOSIUM 9.00–17.30	POLICY SYMPOSIUM 9.00–12.15			
			TECHNICAL SYMPOSIUM 13.45–17.30	TECHNICAL SYMPOSIUM 9.00–17.30	TECHNICAL SYMPOSIUM 9.00–17.30	TECHNICAL SYMPOSIUM 9.00–17.30
			ECONOMIC SYMPOSIUM 9.00–18.00	ECONOMIC SYMPOSIUM 9.00–18.00	ECONOMIC SYMPOSIUM 9.00–18.00	

DIGITAL MICROWAVE RADIO SYSTEM SAMI-15

Official permission No: MR 96/91

The digital microwave equipment family SAMI has been developed for transmitting PCM telephone channels and data transmission in computer networks. Its versions are operating in the 1,5...23 GHz frequency range as reliable telecommunication links in public or private networks on short and medium-long distances. The range can be increased by repeater stations.

The equipment group type SAMI-15/P,S is utilized at 15 GHz for transmission of 2,048 and 8,448 or 4x2,048 Mbps rate signal flows in systems with or whithout stand-by. PRIMARY OR SECONDARY PCM MULTIPLEX SIGNALS (30/120 telephone channels), OR SIGNALS OF COMPUTER NETWORKS CAN BE FORWARDED. Further functions: SERVICE TELEPHONE, REMOTE CONTROL, AUXILIARY DATA CHANNEL.
A radio equipment group consists of an outdoor and indoor unit, interconnected by a max. 300 m long complex cable. The OUTDOOR UNIT is a microshelter (containing the RF assemblies and the signal handling circuits) built together with the antenna. Parts of the INDOOR UNIT are: digital connecting circuits, service channel assemblies, and — optionally — stand-by circuits and secondary multiplex circuits. On order: installation, putting into operation, etc., on demand: other performances — e.g.: tv+4 sound channels — are possible.

PLEASE CONTACT US FOR DETAILED INFORMATION!

TKI®

RESEARCH INSTITUTE FOR TELECOMMUNICATIONS
TÁVKÖZLÉSI KUTATÓ INTÉZET
ADDRESS: 1525 BUDAPEST, P.O.B. 15

Tel: 135-3900
Fax: 135-5560
Telex: 22-4338

THE BUSINESS, OFFICIAL OR PRIVATE INFROMATION SHOULD BE PROTECTED AGAINST GETTING INTO INCOMPETENT HANDS. The open message forwarding is not a secure way, since it enables anybody to catch the message by using some simple trickery. This is also especially dangerous in case of picture information, carrying business secrets, signatures, banking data, etc.

THE QUICK AND SAFE GUARDING CONTRIVANCE
OF TELEFAX MESSAGES IS THE

faxGUARD TYPE P100f

BEING A TELEFAX CYPHERING UNIT INTERFACED TO ONE OF MOST WIDESPREAD TELEFAX DEVICES TYPE CANON 270S. The faxGUARD has the permission No: 213-04770.

Primary users of faxGUARDs are: banks, banking houses, insurance companies, state-owned and private firms, autonomies, administrative authorities, political and social organizations.

THE faxGUARD OFFERS THE FOLLOWING SERVICES: • open or cyphered telecopy of documents • automatic recognition of the open or cyphered mode • no limitation of the basic services of the telefax unit • cryptographic algorithms: DES; key-dimension: 56 bits, or HES; key-dimension: 126 bits • operation is served by menu systems • establishing of network of hierarchical access, • three-level key system, chip-card entering.