Generalized Automatic Phonetic Transcribing of Speech Signals

Taras K. Vintsiuk

UNESCO/IIP International Research-Training Centre for Information Technologies and Systems,
40 Academician Hlushkov Avenue, Kyjiv 252022 Ukraine
Тел.: +380 44 266-4356
vintsiuk@uasoiro.freenet.kiev.ua

ABSTRACT

A so-called generalised phoneme recognition problem for the two-level speech understanding system is being solved. It means that under free phoneme order it is being found the $N>>1$ best phoneme sequence recognition responses. The method is based on constructive description of diverse realisations of a speech signal. A stochastic generative automata grammar, which is assigned to synthesise the speech signal prototypes, serves for it. This grammar composes all possible speech signal prototypes with allowance for non-linear rate of pronouncing in general, and of the pronouncing of individual phonemes in particular, as well as co-articulation and reduction of sounds and non-linear variation of the speech signal intensity along the time axis. To make deeper the earlier fulfilled research, phoneme-threephones (PT) signal prototypes are introduced. Rules for joining of PT signal prototypes into sequences are evident: the output and input phonemes of joining PT have to coincide. The problem is being solved using new compu-tational scheme of dynamic programming, based on (for substantial reduction in both memory and calculation requirements) the concepts of potentially optimal index and phoneme response.

АБСТРАКТ

Тарас Вінцюк. Узагальнене автоматичне фонетичне транскрибування усномовного сигналу. Розв'язується проблема так званого узагальненого автоматичного транскрибування усномовного сигналу, яка виникає при створенні дворівневої системи розміння мови. Вона полягає у знаходженні $N>>1$ найкращих послідовностей фонем, які складають відповідь розпізнавання. Метод грунтується на конструктивному описі (заданні) всього розмаїття мовних сигналів. Для цього використовуються стохастичні автоматні породжувальні граматики, які синтезують модельні ("еталонні") сигнали зв'язної мови, що відрізняються нелінійно змінюваними в часі темпом та інтенсивністю вимовляння, враховують коартикуляцію та редукцію звуків, індивідуальні особливості голосу. Щоб більш адекватно врахувати змінюваність мовних сигналів, введені поняття фонем-трифонів та їх модельних сигналів, індивідуального усномовного файлу (паспорта). Правила об'єднання модельних сигналів фонем-трифонів в послідовності є очевидними: вихідне ім'я та вхідне ім'я двох сусідніх фонем-трифонів повинні збігатись. Проблема узагальненого автоматичного фонетичного транскрибування розв'язується за допомогою ефективної процедури динамічного програмування, в якій, з метою значного скорочення обсягів обчислень та пам'яті, використані поняття потенційно-оптимальних індексів та потенційно-оптимальних фонемних відповідей розпізнавання.

INTRODUCTION

Still it is retained popular such approach in automatic speech recognition and understanding. It assumes that firstly continuous speech must be recognised as phoneme sequence, and then this phoneme sequence must be recognised and understood as word sequence and meaning to be transmitted by a speech signal [1, 2].

Though this approach seems to be erroneous, since the best method of finding of phonemes to be transmitted is both to recognise and to understand a speech signal, however it shows a preference for simplifying the research job distribution between specialists in acoustics, phonetics, linguistics, informatics.

To get better this approach it was proposed to introduce significant decisions in phoneme recognition procedures [2, 3]. The next step consists in making improvements to used generative automata grammars, for example instead of phoneme-diphones speech model [1, 3] to put into operation a phoneme-threephones one.

In this paper it is proposed a so-called generalised phoneme-threephone recognition problem for the two-level speech understanding system. The structure of this system is shown in Fig. 1. A generalised phoneme recognition problem means that under free phoneme order it is being found the $N>>1$ best phoneme sequence recognition responses. Then a Speech Interpreter analyses these phoneme sequences through Natural Language Knowledge filter.

PHONEME RECOGNITION IN CONTINUOUS SPEECH. GENERAL IDEA

The general idea is, taking into account inertial properties of articulation apparatus and language phonetics only, to construct some PT generative automata grammar which can synthesise all possible continuous speech model signals (prototypes) for any phoneme sequence. This grammar has to reflect such phenomena of speech signal variety as non-linear change of pronouncing both rate and intensity, sound co-articulation and reduction, sound duration statistics, phonemeness, and so on. Then the phoneme-by-phoneme recognition of unknown continuous speech signal will be involved in a synthesis of the most likely speech model signal and a determination of the phoneme structure of the latter.

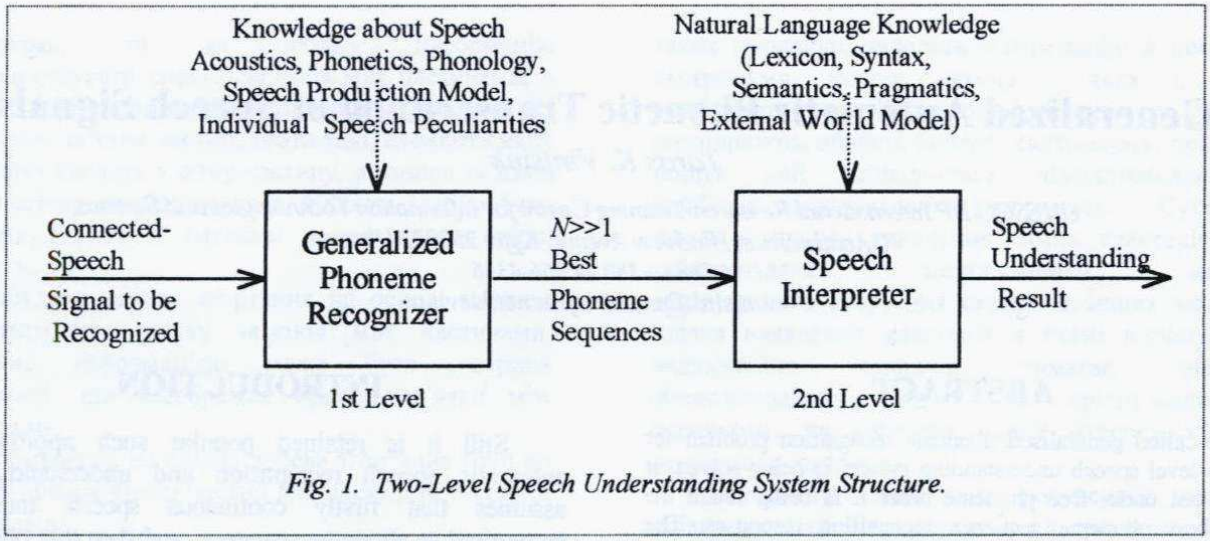

Fig. 1. Two-Level Speech Understanding System Structure.

The problem of directed synthesis, sorting out and formation of a phoneme sequence recognition response is solved by using a new computational scheme of dynamic programming, in which (for a substantional reduction in memory and calculation requirements) the concepts of potentially optimal both index and phoneme are used.

At first the phoneme-by-phoneme continuous speech recognition problem will be considered. Then this statement will be generalised for $N>>1$ best phoneme sequences.

GENERAL FREE PHONEME-THREEPHONE SEQUENCES GENERATIVE GRAMMAR

This mentioned generative grammar for free phoneme sequences will be given under PT interpretation.

Let be given the finite set K of the phonemes $k \in K$. The phoneme alphabet includes the phoneme-pause #. In K there will be distinguished stressed and non-stressed vowels, hard and soft consonants, stationary phonemes like $k \in$ {A, O, U, E, I, Y [all stressed and non-stressed], V, V' [the symbol ' denotes soft-ness], ZH, Z, Z', J, L, L', M, M', N, N', R, R', F, F', KH, KH', SH, #} $\equiv K^{st} \subset K$, which change their duration, and transitive phonemes $k \in$ {B, B', G, G', D, D', K, K', P, P', T, T'} $\equiv K^{tr} \subset K$.

Then there are considered all possible PT or about 2,000—3,000 basic PTs $t \in T$. Each PT t from the PT alphabet T besides the name t has also the triple name $t=uWv$ where $u,W,v \in K$ and u,v are input and output phoneme names for PT t, respectively. So the PT $t=uWv$ is the phoneme W that is considered under influence of neighbouring phonemes u and v in context, they are the first u which precedes W and the second v which follows W.

From now on we will assume that besides phoneme and PT alphabets there are given such knowledge:

A. A finite set E of elementary speech signal prototypes or typical one-quasiperiodical segments $e(j) \in E$ where $j \in J$ is a $e(j)$ name in the name alphabet J. E.g. there are $|J| = |E| = 2^{16}$ elements in E and J. So the set J makes the microphoneme level of speech patterns and the pair (J, E) is the code book for one-quasiperiods.

B. A finite set T of PT $t \in T$. The PT t is specified by its acoustical transcription in the alphabet J: $t = (j_{t1}, j_{t2}, \ldots, j_{ts}, \ldots, j_{tq(t)})$, where s indicates the ordinal place in the transcription t and $q(t)$ is the transcription duration for t.

C. Distributions $P(x/j)$ of observed elements (quasiperiods) x for all $j \in J$, particularly $P(x/j)=P(x/e(j))$.

The knowledge mentioned in A, B and C are found at training mode [1, 2]. For each speaker they form a so-called Speech Speaker file.

After the preprocessing a speech signal to be recognised is presented by the sequence X_{ol} of observed one-quasiperiodical segments or elements x_i: $X_{ol} = (x_1, x_2, \ldots, x_i, \ldots, x_l)$, where l is the quantity of observed quasiperiods. The segment $X_{mn} = (x_{m+1}, x_{m+2}, \ldots, x_n)$, $0 \le m < n \le l$ is considered as a signal realisation of the PT t with the probability which is calculated as the convolution on microphonemes bounds $\{r_s\}$:

$$P(X_{mn} / t) = \max_{\{r_s\}} \prod_{s=1}^{q(t)} \prod_{i=r_{s-1}+1}^{r_s} P(x_i / j_{ts}), \qquad (1)$$

where $r_0 = m$, $r_{s-1} < r_s$, $r_{q(t)} = n$. The respective stochastic generative automata grammar (graph) for both PT model signals generating and comparison of the signal segment X_{mn} with all generated ones accordingly to (1) is shown in Fig. 2a. That graph has $q(t)$ states. To each state s it is ascribed the microphoneme $j(s) = j_{ts}$ with the distribution $P(x/j_{ts})$. The transitions between states are doing in accordance to arrows and during 0 or 1 discrete time steps. It is forbidden to remove microphonemes here. The grammar shown in Fig. 2b forbids to remove more than two microphonemes running. Schematic notes for PT graph $t=uWv$, $u,W,v \in K$ are given in Fig. 2c, where only the input $s=u$ and the output $s=v$ states are distinguished.

Let us unite all PT graphs into common one. It is permissible to connect PT into phoneme sequences so

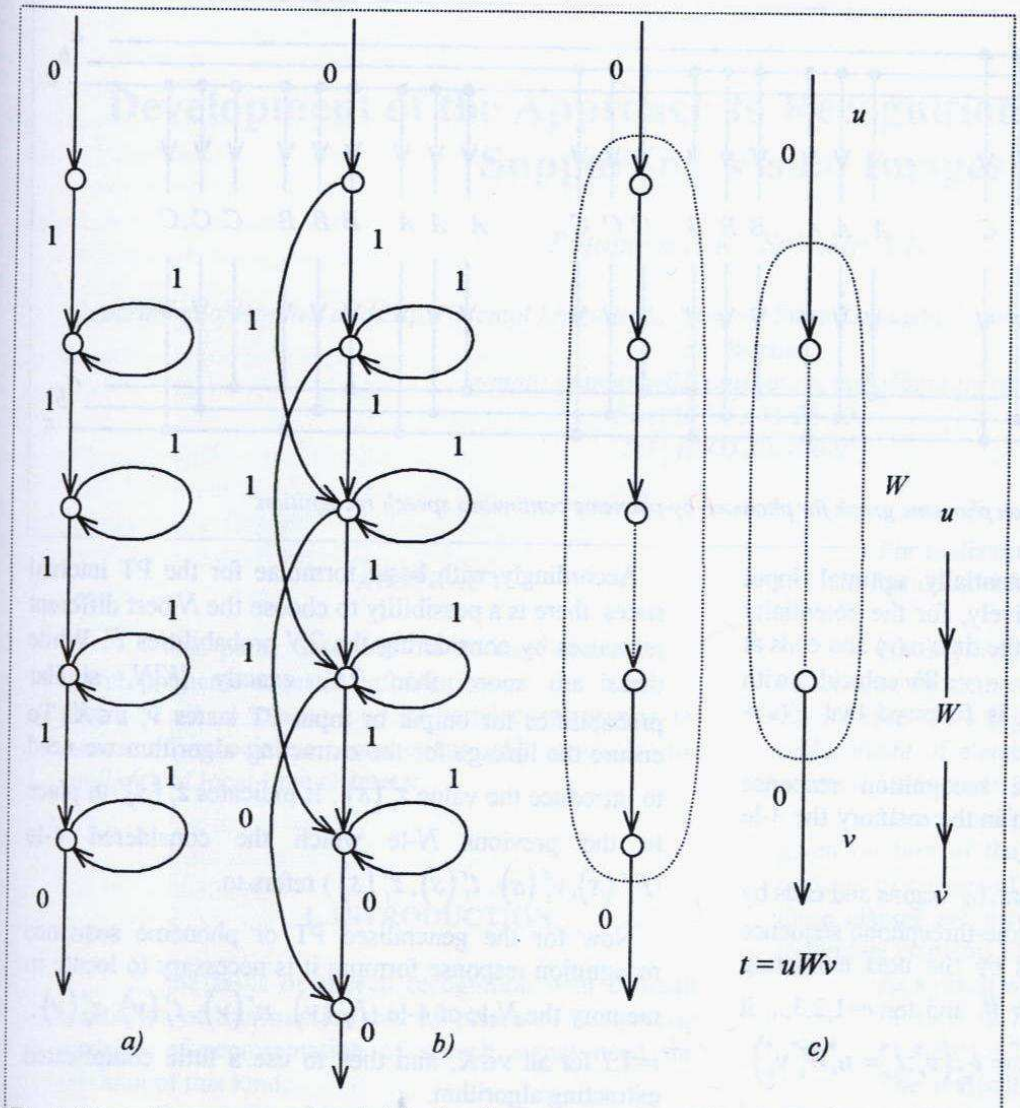

Fig. 2. Generative grammars (graphs) for the phoneme-threephone: a) no microelement omis-sion; b) no two microelements running omission; 3) schematic notes of the PT graph t = uWv.

that the output phoneme name of preceding PT coincides with the input phoneme name of following one. It means that the input and output states for different but permissible for linking PT have to coincide.

Going such a way it will be received the common phoneme graph (CPG) for continuous speech signal generation. This full CPG for three phoneme alphabet $K=\{a, b, c\}$ is shown in Fig. 3. It is distinguished the input state $s=u$, the output one $s=v$ and internal states s for each PT $t=uWv$, $u,W,v\in K$. One of the states $s=k\in K$ in Fig. 3 is associated with the phoneme-pause #. Let us introduce the overall enumeration of states in the CPG accordingly with a permissible movement along the arrows.

Looking into CPG the best phoneme sequence recognition response or, that is the same, the best permissible PT sequence recognition response is defined by maximisation of the expression (2):

$$P\left(X_{0l}/\left(t_1,\ldots,t_s,\ldots,t_Q\right)\right)=\max_{\{r_s\}}\prod_{s=1}^{Q}P\left(X_{r_{s-1}r_s}/t_s\right), \quad (2)$$

where $\{r_s\}$ are the bounds between phonemes-threephones in X_{0l}.

PHONEME SEQUENCE RECOGNITION ALGORITHM

Let be designated by $\Omega_i(s)$ a set of continuous speech prototypes of duration i which are generated by the CPG as a result of movement from state $s=\#$ to state s within i time steps. Let be denoted by $F_i(s)$ the best probability (2) which is reached on the set $\Omega_i(s)$ but for the initial segment $X_{0i}=(x_1, x_2,..., x_i)$, and by $n_i(s)$ the potentially optimal beginning of the last PT $t_i(s)$ in the best PT sequence for $\Omega_i(s)$.

Let $F_r(s)$, $n_r(s)$, $t_r(s)$ have been calculated for all states s and for all time steps $r<i$ which precede i. Then after the next observed element x_i appearance simultaneously (in parallel) for all states s new values $F_i(s)$, $n_i(s)$, $t_i(s)$ are calculated in order a), b), c):

a) for all internal PT states $s\in t=uWv$, besides PT first states, and for all t (see Fig. 2a):

$$F_i(s)=\max\{F_{i-1}(s\text{-}1),\ F_{i-1}(s)\}\cdot P(x_i/j(s)),$$

$$n_i(s)=\begin{cases}n_{i-1}(s-1), & if\ \ F_{i-1}(s-1)\geq F_{i-1}(s);\\ n_{i-1}(s), & if\ \ F_{i-1}(s-1)<F_{i-1}(s);\end{cases}$$

b) for all first internal states $s=s_1(t)\in t=uWv$ and for all $t\in T$ (see Fig. 2a):

$$F_i(s_1(t))=\max\{F_{i-1}(u(t), F_{i-1}(s_1(t))\}\cdot P(x_i/j(s_1(t))),$$

$$n_i\left(s_1(t)\right)=\begin{cases}i-1, & if\ \ F_{i-1}\left(u(t)\right)\geq F_{i-1}\left(s_1(t)\right);\\ n_{i-1}\left(s_1(t)\right), & if\ \ F_{i-1}\left(u(t)\right)<F_{i-1}\left(s_1(t)\right);\end{cases}$$

c) for all common output states v of all PT $t=uWv$ with the same $v(t)=v$:

$$F_i(v)=\max_{t=uWv:v(t)=v}F_i\left(v(t)\right),$$

$$t_i(v)=u_i(v)W_i(v)v=\arg\max_{t=uWv:v(t)=v}F_i\left(v(t)\right),$$

$$n_i(v)=n_i\left(v\in t_i(v)=u_i(t)W_i(t)v\right),$$

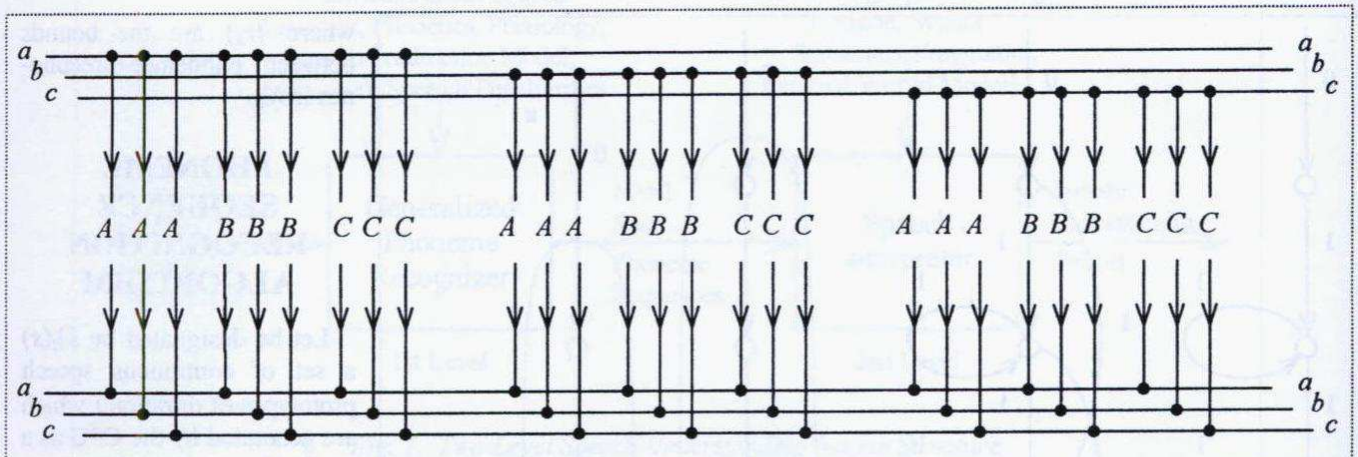

Fig. 3. Common phoneme graph for phoneme-by-phoneme continuous speech recognition.

where $u_i(v)$ and $W_i(v)$ are potentially optimal input phoneme and phoneme, respectively, for the potentially optimal PT $t_i(v)$ which begins at the time $n_i(v)$ and ends at the time i ; since each input state $u \in t=uWv$ coincides with the same output state u then it is followed that $F_i(u)=F_i(v=u)$.

For the phoneme sequence recognition response forming it is sufficient to remain in the memory the 3-le array $F_i(v)$, $n_i(v)$, $t_i(v)$, $i=1:l$, $v \in K$.

Since a continuous speech signal X_{0l} begins and ends by a full PT $t=\#\#\#$ then the phoneme-threephone sequence recognition response is formed by the next extracting algorithm. Let be $n_1^* = l$, $t_1^* = \#$ and for $r=1,2,3,...$ it will be extracted $t_{r+1}^* = t_{n_r^*}\left(v_r^* : t_r^* = u_r^* W_r^* v_r^*\right)$, $n_{r+1}^* = n_{n_r^*}\left(v_r^* : t_r^* = u_r^* W_r^* v_r^*\right)$ until $n_{r+1}^* = 0$ will be reached. Then the PT sequence t_r^*, $r=1,2,3,...$ will be the PT recognition response in the opposite direction and n_r^*, $r=1,2,3,...$ will be the respective PT bounds in the signal X_{0l}. The phoneme sequence recognition response will be W_r^*, $r=1,2,3,...$.

To begin the recognition process it is assigned $F_0(u=\#)=1$ and $F_0(s)=0$ for all other states $s \neq 0$.

THE GENERALISED ALGORITHM

To find $N>>1$ best phoneme or PT sequences in the signal X_{0l} let us modify the basic algorithm.

Now for all states s in the CPG and for any time step i it will be calculated N-le of not 3-le but 4-le $(F_i^r(s), n_i^r(s), t_i^r(s), z_i^r(s))$, $r=1:N$ which is composed of N best probabilities $F_i^r(s)$ that correspond to N best but different PT sequence recognition responses for X_{0i}..

Accordingly with basic formulae for the PT internal states, there is a possibility to choose the N best different responses by considering the $2N$ probabilities F. While there are more than $2N$, exactly $|K|N$, similar probabilities for output or input PT states v, $u \in K$. To ensure the linkage for the extracting algorithm we need to introduce the value $z_i^r(s)$. It indicates $z_i^r(s)$ -th place in the previous N-le which the considered 4-le $(F_i^r(s), v_i^r(s), t_i^r(s), z_i^r(s))$ refers to.

Now for the generalised PT or phoneme sequence recognition response forming it is necessary to locate in memory the N-le of 4-le $(F_i^r(v), n_i^r(v), t_i^r(v), z_i^r(v)$, $i=1:l$ for all $v \in K$, and then to use a little complicated extracting algorithm.

CONCLUSION

There exists such an opinion that it is possible to design a machine for automatic phoneme recognition in continuous speech without any appealing to speech understanding. Here it is proposed one effective robust algorithm for this problem solving which guarantees $N>>1$ best phoneme sequence responses finding.

ЛІТЕРАТУРА

1. Vintsiuk T.K. , *Avtomatyka* 6, 40 - 49 (1972); 1, 63 - 72 (1973).
2. Vintsiuk T.K., *Analysis, Recognition and Understanding of Speech Signals*, Kiev: Naukova dumka, 1987, 264 p.
3. Vintsiuk T.K. , "Generelized Problem for Automatic Phoneme Recognition", *Proceedings of the Workshop SPECOM'97*, Cluj-Napoca, Romania, pp 115 - 118, 1997.

Development of the Approach to Recognition of Speech With a Support on Visual Images

Potapova R.K., Sobakin A.N.

Department of Applied and Experimental Linguistics, Moscow State Linguistic University, 125445, Moscow, Ostozenca, 38, Russia
e-mail: potapova@linguanet.ru; mglu@online.ru
Fax: (095) 246-28-07;
Tel.: (095) 201-56-97

ABSTRACT

This paper describes the concept of receipt of useful phoneme-acoustic information on the bases of nature of visual images (and its parts) configuration by means of associative connections [1] specially for similarity of local-time character.

1. INTRODUCTION

The tasks of speech recognition with difficult conditions of transmission and by means of not precise methods of representation of speech signal need the research of this kind.

Except for a hierarchical way of the account of feature units the parallel analysis of the quantitative characteristics of offered attributes is possible. **The method** realizing parallel way of classification of words, is based on use of a matrix of affinity received for the relation of belonging to classes of words. The given descriptions of two words concern to one class, if taxonomic distance between the descriptions in space is least.

2. METHOD AND EXPERIMENT

With the purpose of demonstration of serviceability of an offered technique we shall consider procedure of classification of the descriptions of ten words (for figures) above mentioned. As the initial data we use values of related features for them [2].

In connection with that a number of features used for the description of the images of words, is given by several values minimal and maximal, for several figures of the image of a word etc.), with the purpose of simplification of the further calculations we shall take advantage only of their average values.

According to the offered attributes we shall receive a matrix of distances (tab. 1).

For realization of procedure of classification we use the feature of a belonging for threshold value equal 10.

Having applied the chosen threshold value concerning all elements of a matrix of distances (see tab. 2), we received a logic matrix of the given relation, by replacement of elements exceeding 10: "zeros", and all others – "ones" (tab. 2).

Having analysed a logic matrix, we see, that the given relation of the belonging are derivated by some variants of classes of equivalence (groups) of words. In these classes are included words, to which correspond individual elements of a matrix of the relation.

As a result we received:
{1.4,8}, {2,7,9,0}, {3,5}, {6} or
{1.4,8,0}, {2.7,9,}, {3}, {5.6}.

The multialternativeness of classification is a consequence of intransity of logic matrix of the given relation. It is easy to be convinced of it, having applied to it criterion of check of property of transitivity. It is necessary to apply to elimination of multialternativeness a method of transitive short circuit consisting in multiplication of a logic matrix on themselves so long as its elements will cease to change.

Division of initial set of the images of words into groups more large, than the separate word, is received by virtue of that circumstance, that we used in example only four features from seven, used at direct classification.

Received preliminary results of use offered in the paper attributes of the approached description of speech images of separate words confirm an opportunity of their division with the help of this system attributes.

At the same time it is necessary to note preliminary character of such conclusion owing to limited volume of sample of images used for experiment both by amount of realizations, and on number of the speakers.

For a substantiation of more universal character offered or not how many modified system of features will be carried out more extensive experiment: the experimental and number of the subjects increased.

Besides the preliminary conclusion about an opportunity of development of the automated technique of classification of the images of words and its realization can be made on the basis of a parallel way of classification.

Processing of parametrical representation of words received with use of band-spectrograph can be considered in quality of one of tasks of statistical processing of the graphic information having two levels of "brightness" in each "point" of the image.

Each realization of a word of one speaker in such graphic representation has its own frame, the beginning and the ending of this one is defined accordingly by beginning and ending of pronouncing of the given speech formation.

The beginning of a word in all experiments was determined on occurrence of the first not "zero" in one of frequency channels from F1 up to F2 (in frequency interval it corresponds with frequencies from 100 Hz up to 4000 Hz). Not "zero" values of each of the specified ranges correspond to excess of accumulation for 20 ms of energy of a speech signal of some threshold value.

Thus, size of threshold value, its choice in relation to level of a signal are very important for steady and reliable finding of the left border of the frame. From the further description of a way of statistical processing of frame it becomes clear, as far as this parameter is important for all procedure of statistical recognition in whole.

The right border of frame of the image of a word was determined similar on last value distinct from zero, of a level of a signal, in same frequency ranges also depend on the same threshold value.

Each realization of a word frame was put in conformity the contrast (graphic) image, on the basis of which statistically were created model sequences.

Creation of model sequences: for creation of model sequences each frame with image of a word was exposed to preliminary transformation which consist in the following.

The image was transformed into a sequence of value of brightness (zero and one) by means of transformation of everyone vertical frequency section of a speech signal in a horizontal vector. Such transformation was carried out by "turn" of a vertical vector on 90 degrees clockwise, that transformed a vertical vector in horizontal. Its left values corresponded with the bottom frequency ranges, and right – high frequency ranges. Each subsequent vector incorporated with previous on the right and, thus, long horizontal vector – was formed line (sequence from "zero" and "ones").

After the described preliminary transformation of each frame of the image we have a set of sequences for set of pronouncings of ten words by ten speakers. This set of sequences serves a statistical material for formation of model sequences.

Formation of a model sequence for each word (class) was carried out by calculation for each position of a sequence of probability of occurrence in it of unit on ten speaker realizations. In practice not probability of occurrence of unit, and size appropriate to numerator of this probability was used. In other words, was counted up amount of units in each position of sequences for the various speakers.

Sequence, received on the basis of statistics, was accepted for reference for the given word (class).

Reference sequences for everyone thus were created words (class). In the given experiment number of classes of recognition coincided with amount of words of recognition and was equaled to ten. As much was created of reference sequences.

At a stage of recognition the showed(presented) sample of a word as a sequence of "zero" and "ones" is compared to the standards and the measure of similarity " of a represented sample for each standard is calculated ". The algorithm of calculation " measures of similarity " creates in set of binary sequences metric space being base for each algorithm of races cognition.

In considered space of binary sequences the metrics as the sum of conditional probabilities of occurrence of "one" or "zero" is offered in each position of a sequence.

This sum of conditional probabilities in view of the made remarks is counted up as follows. If in the presented standard in some position there is a "one", in the sum the number of "ones" from reference is added from sequences of a checked class in the same position. If in the presented standard in a considered position there is a "zero", in the sum the difference is added. Last action corresponds to addition in the sum of probability of occurrence of "zero" (additional probability) in the given position.

The complete summation is made on length of the presented word. In the volume a case, when the word comes to an end before the standard (model), procedure of summation comes to an end. If the standard is shorter than a word, it is supplemented in "zero", that corresponds to "zero" probability of occurrence of "ones" in these positions.

The received sum is normalized on the length of presented word, and this sum is accepted for a measure of similarity of the presented word.

The greater value of a measure of similarity corresponds to greater "similarity" of the presented word to the given class. The maximal value of measures of similarity on all classes will correspond hypothetically to belonging of showed word to this class.

3. CONCLUSION

The described algorithm of recognition was examined on a material to ten words realized by ten speakers.

Experimental results of comparison of ten words pronounced by the speaker C., with reference sequences are given in the table 3. It is necessary to note,

that the showed speech samples of the speaker C. not participated in reception of the samples for comparison.

As it is visible from the given table, the greatest value of a measure of similarity are located on the main diagonal of a matrix. It means in the whole serviceability of such approach in recognition of the images of words.

At the same time, there are separate difficult cases of comparison for offered measures of similarity. In particular, the measure of similarity of a word "eight" with various sample sequences has no the brightly expressed maximal value, that can be explained as large variability of separate speaker realizations of the word which has served with a basis for creation of the sample of the appropriate class ("eight").

Variability of speaker realizations in this class is explained, on the one hand, by various degree of a reduction of past-stressed vowel, on the another – by various degree of coarticulation, penetrating all word as a whole.

The possible ways of increase of reliability of algorithm of recognition consist in expansion of statistical base of creation of the samples and improvement of techniques of comparison of speaker realizations with the samples.

REFERENCES

1. R. K. Potapova. "Ob odnom podkhode k klassifikatsiji reche-zritelnych obrazov na baze assoziativnykh svjazej. Mater. VI Vseros. Konf. "Nejrokompjutery i ikh primenenije". M., 2000.
2. R.K. Potapova. "Rech: kommunikatsija, informatsija, kibernetika." Radio i Svjax. M., 1997.

APPENDIX

Table 1. A matrix of distances

Word	1.	2.	3.	4.	5.	6.	7.	8.	9.	0
1.	0	11,25	24,29	1,02	21,4	16,05	14,09	7,28	12,17	6,20
2.		0	21,26	12,03	16,28	10,68	3,61	13,91	9,30	8,21
3.			0	23,85	6,64	11,05	21,01	18,5	13,18	18,00
4.				0	21,77	16,08	14,86	7,24	12,50	6,65
5.					0	6,35	15,34	17,85	9,73	15,24
6.						0	10,45	12,88	4,95	9,46
7.							0	16,43	9,63	10,35
8.								0	10,60	6,44
9.									0	6,06
0										0

Table 2. A logic matrix of distances

Word	I	2	3	4	5	6	7	8	9	0
I	1	0	0	I	0	0	0	1	0	I
2		1	0	0	0	0	I	0	I	I
3			1	0	I	0	0	0	0	0
4				I	0	0	0	I	0	I
5					I	I	0	0	I	0
6						I	0	0	I	I
7							I	I	0	I
8									0	I
9									I	I
0										I

Table 3. Experimental results of comparison of ten words, pronounced by the speaker C., with reference sequences.

Word	"0"	"1"	"2"	"3"	"4"	"5"	"6"	"7"	"8"	"9"
"0"	6,11	4	4,69	3,7	5,37	3,96	4,18	3,35	5,02	4,99
"1"	4,43	6,54	3,25	2,63	5,65	3,12	2,94	2,05	4,31	4,1
"2"	5,28	3,07	6,77	4,55	3,48	5,73	4,84	3,79	4,11	5,08
"3"	2,84	3,3	3,81	7,84	4,08	5,35	5,47	5,93	4,53	5,65
"4"	4,41	5,28	3,69	4,5	7,51	3,77	4,44	3,45	4,74	5,74
"5"	3,98	3,47	4,37	5,34	3,32	7,15	4,32	4,26	3,10	5,01
"6"	4,51	4,57	5,46	6,7	6,4	5,77	7,54	5,52	5,8	6,21
"7"	2,73	2,89	3,41	6,63	3,86	4,65	3,91	7,06	3,48	5,02
"8"	4,98	4,86	4,04	4,76	5,62	4,25	4,47	3,52	6,06	6,05
"9"	4,31	3,31	4,48	5,59	5,96	4,89	5,93	4,73	5,42	7,43

IDENTIFYING TRANSIENT PATTERNS IN IDIOMATIC GREEK AND SLAVONIC MUSIC

*Dionysios Politis** *Alexander Dolia*† *Panagiotis Linardis***

{*Multimedia Lab, **VLSI Lab},
Computer Science Dept., Aristotle University of Thessaloniki, GR-540 06, GREECE.
e-mail: {dpolitis, linardis}@csd.auth.gr
† Dept.504, State Aerospace University,
17 Chkalova St., Kharkov, 61070, Ukraine.
e-mail: lukin@mmds.kharkov.ua

Abstract—**This paper describes a method for determining the content affinity of musical patterns in vocal reproduction in Greek and Slavonic by using wavelet decomposition techniques. The goal of the analysis is to determine the exact musical description of the patterns in order to reproduce them by musical synthesizers acting as synthetic singers. The exemplar pattern of this paper is that of "petasti", which was selected because it is a transient phenomenon particular in traditional Greek songs and chants, in both instrumental and vocal performances. A comparison of the pattern is attempted with Slavonic chants by determining time and frequency localizations of the pattern.**

Index Terms—**Wavelet decomposition, musical patterns, "petasti", voice synthesis.**

I. INTRODUCTION

This paper copes with the identification of musical patterns that literally exist in Greek and Slavonic musical traditions. The origin of these patterns can be found in manuscripts that are at least 10 centuries old. In Fig. 1.(a) a fragment of a manuscript is shown which implicitly denotes these patterns [1]. In Fig. 1.(b) explicit marks of "petasti" are recorded in Byzantine music notation [2].

The creation of a musical notation system proved to be a painful and long procedure in the West and especially in the East, where as the manuscripts that have survived indicate (Fig. 1), we have had different notations evolving one from the other in a shorthand like form. This musical system was not confined only to ecclesiastical music; it was a generalized musical system originating directly from ancient Greek Music and was used as the usual music surface by all the people living in the vast areas of the Byzantine empire, from Southern Italy and the Balkans up to Ukraine and Russia and down to Middle East and Egypt.

Finally the Byzantine notation prevailed in the East as the dominant musical notation and was reformed in 1814 and 1881 from committees of the Ecumenical Patriarchate of Constaninople to an analytical notation

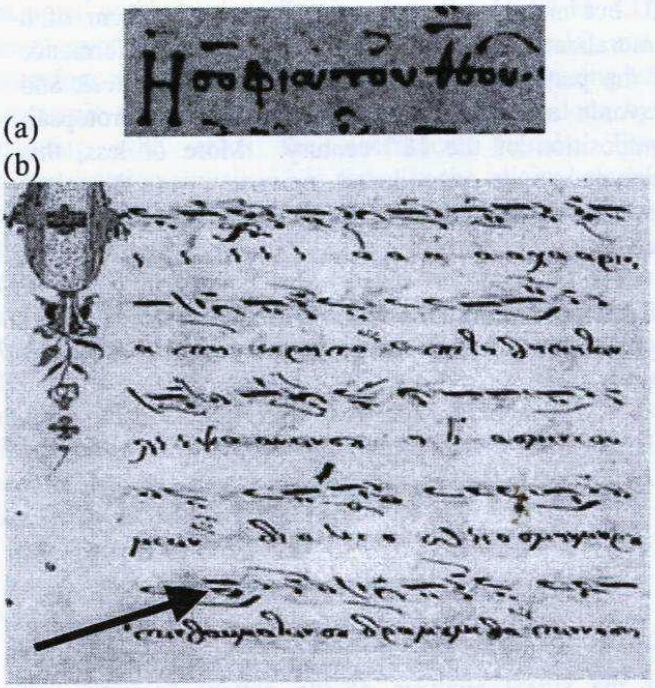

Fig. 1. Byzantine Music Manuscripts: (a) Detail from an early manuscript, the so-called 'Chartres fragment', with musical notation, beginning of the sticheron "Η σοφια του Θεου, Mode plagal D", Copenhagen. (b) "Petasti" pointed out in a Doxastarion composed by Jacov Protopsalt, Mount Athos, 1805 AD.

system whose symbols came out of the numerous symbols of the earlier shorthand-like notations [3].

Major differences between the polyphonic Western music tradition and the Byzantine system are:

- the intervalistic system (system of musical scales and their microtonal distribution) that contains predominantly musical intervals that are smaller than the well tempered ones. Successors of the ancient Greek Dorian, Lydian, Phrygian, Mixolydian Modes and their plagal ones are encountered [3][4]. This characteristic introduces a subjective criterion of orientalization in the psycho-acoustics of the hearings, and
- the use of larger, formalized transitory patterns as main elements of the musical structure that are

inherent in vocal performance and render a qualitative nuance.

A musical pattern that has a quantitative and qualitative as well character is that of "petasti". This pattern has been detected in instrumental performance of traditional Greek songs [5] and in Byzantine music vocal performance [6]. According to Byzantine Music theory [2], "petasti" is a transient phenomenon that assigns a phonation quality to the raising or lowering of pitch. In this paper, analysis of this pattern is performed in order to decipher its qualitative nature. By the term qualitative we mean the characteristics that enable the specialized listener to recognize the "petasti" pattern not simply as a pitch fluctuation, a fluctuation of fundamental frequency F0, but mainly as a parametric transitory form of a gutturalization. Analysis is focused on the performance of the pattern in Mode A (Dorian) [4] in Greek and Slavonic according to the classical Jacov Protopsalt composition of the 18th century. More or less, this performance is reproduced nowadays in the slow rendition of the vesper hymn "Lord, I have cried unto thee".

The exact description of this pattern in terms of microtonal distributions will enable the correct musical reproduction by computer programs acting as synthetic singers.

II. EXTRACTION OF AUDIO FEATURES USING THE DISCRETE TIME WAVELET TRANSFORM

For the signal analysis of the pattern "petasti" the Discrete Time Wavelet transform (DTW) will be used in conjunction with the classical Short Time Fourier Transform (STFT). One major advantage afforded by wavelets is the ability to perform local analysis. This is useful near the discontinuity areas of the signal, namely the consonants for vocal signals. A second advantage is that with each stage DWT the signal is seperated into *approximations* and *details* while downsampling is performed. The DWT can be seen as an equivalent of a tree-structured multirate filter bank (Fig. 2) [7].

The mother wavelet used for decomposition was db4, the fourth of the Daubechies family wavelets, and the wavelet decomposition tree is presented in Fig. 2, albeit for convenience only two levels are drawn. D(z) is the outcome of a high pass filter and downsampling, thus a detail component, and A(z) is the outcome of a low pass filter and downsampling, thus an approximation component. This procedure produces at each stage DWT coefficients which are denoted in Fig. 2 as $\mathbf{cA_i}$ for the approximation components and as $\mathbf{cD_i}$ for the detail components for the **i**-th level analysis of signal S.

In order to analyze the transient nature of "petasti", in both Greek and Slavonic reproductions of classical

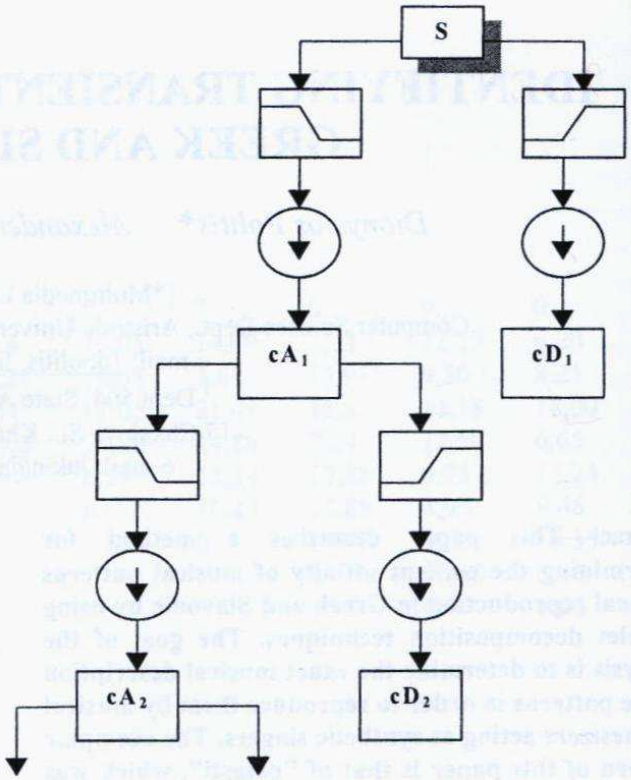

Fig. 2. The 2-level binary tree structured filter bank for signal S.

musical phrases, prototypal performances of these melodies have been recorded.

The waveform of the prototypal performance in Greek is shown in Fig. 3(a) and its spectrogram is shown in Fig. 3(b). In Fig. 3(b) the segment around the first phoneme /i/, the one that is performed with "petasti", fluctuates around note E3 in a manner close to wavelet db4 (Fig. 4). This is the reason for choosing the Daubechies family of wavelets.

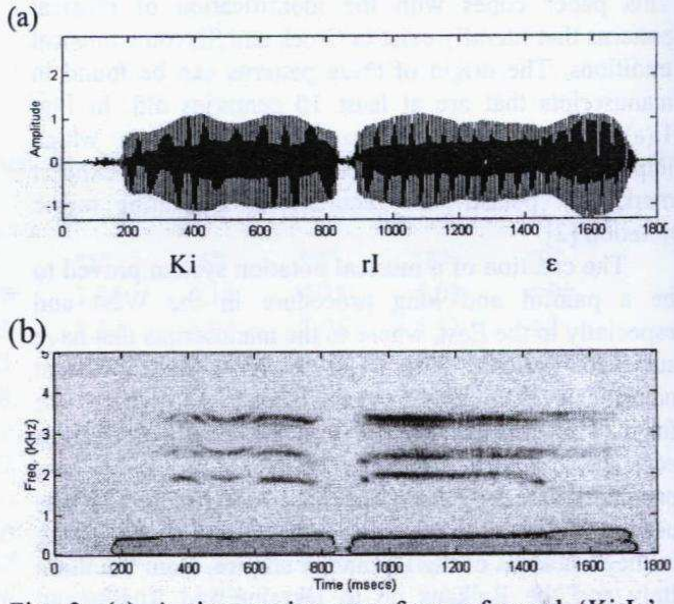

Fig. 3. (a) A time series waveform of word 'KirIε' consisting of three morphemes and a diphone segment. The first phoneme /i/ is performed as a "petasti" pattern. (b) The corresponding spectrogram with emphasized formants F1, F2 and F3.

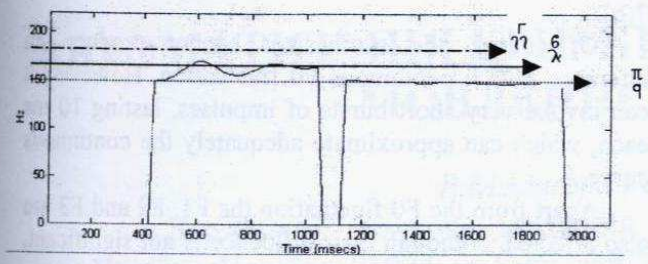

Fig. 4. F0 fluctuations around note Βου, the equivalent of note E3. Note Βου corresponds to 158 Hz, while the pitch levels of notes Πα and Γα, shown with dotted lines, correspond to 144 Hz and 170 Hz respectively. The interval Βου-Γα is slightly bigger than a semitone.

The pattern we examine consorts the first morpheme of the utterance in Fig. 3(a). By using a cepstrum based approach, fundamental frequency F0 is estimated, and the "petasti" fluctuation around note Βου is estimated. Βου is the equivalent of note E3 of the well tempered scale and it is assigned the frequency of 158 Hz [2]. The subject N. Kougias who has performed the musical pattern has achieved a remarkable tuning with the resonant frequencies.

As seen from the formant curves of Fig. 3(b), it is clear that the musical pattern affects not only F0 but also the first three formants, F1, F2 and F3.

The waveform of an equivalent prototypal performance in Slavonic is presented in Fig. 5(a). The word uttered is 'Gospodi' with the first /o/ performed with "petasti". For this utterance, the cepstrum based technique for the determination of F0 yields poor results.

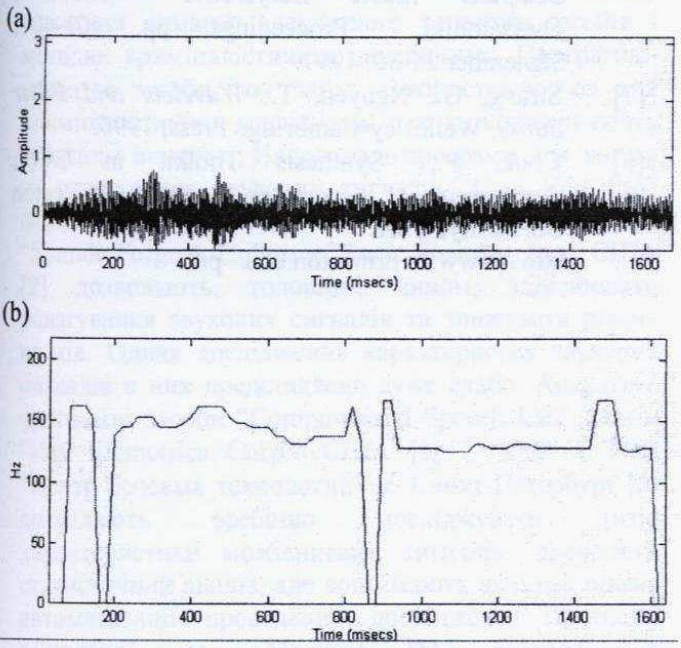

Fig. 5. (a) Time series of morpheme /Go/ from the slow rendition of the word 'Gospodi' having the first /o/ performed with "petasti". (b) F0 cepstrum based estimation of approximation cA3 of the original signal.

In this case, the signal was decomposed by using the fourth wavelet of the Daubechies family and was analyzed using the third-level approximation of it. Prior to estimating F0, upsampling takes place. Then F0 is readily estimated. The final F0 contour is drawn in Fig. 5(b).

Apart from F0 fluctuations, the first three formants of the utterance also have important information. Following the same decomposition scheme, we obtain the detail coefficients of the analyzed signal in order to estimate the first three formants of the utterance.

We focus our analysis on signals cA3 and cD3. cA3 consists of the approximation coefficients downsampled and decomposed from the original signal by the Discrete Wavelet Transform three times. Both signals are thus sampled at 5512.5 Hz (deriving from the 44100 Hz CD level quality recordings of the original signals) and therefore a resampling procedure takes place aiming to upsample them to at least 11025 Hz. This is done by padding with zeros the coefficient signals. The approximation signal cA3 is virtually the original signal that has become less noisy after the successive decompositions, and the high frequency information concerning formants F1, F2 and F3 is filtered out of the signal in the form of the detail coefficients signal cD3.
In Fig. 6 we see an estimation of the formant fluctuation for the utterance with "petasti" of /o/ (Fig. 6(a)). It is evident that in this case the performed analysis is inadequate. Consequently, we dilate the signal estimating its approximation and detail coefficients according to the binary-tree structure of Fig. 2.

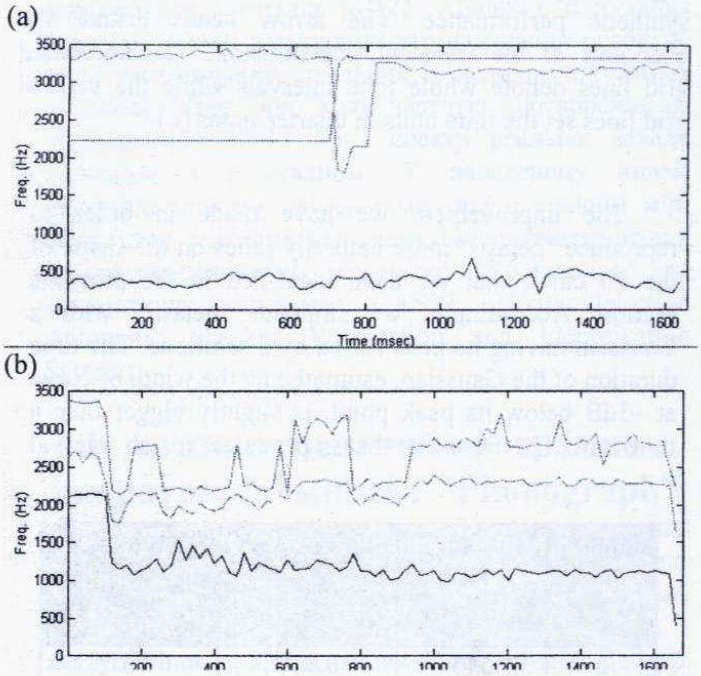

Fig. 6. (a) Formants F1, F2 and F3 estimated from the prototypal signal in Slavonic (b) Formants F1, F2 and F3 estimated from cD3, the detailed third level high frequency component of the signal.

III. CONCLUSIONS: ADVANCES IN SYNTHETIC VOICE REPRODUCTION

The synthetic musical performer we have used to reproduce melodic themes with "petasti" was a formant synthesizer that had physically modeled the glottis and shape files of a Greek singer. This program was compiled using P. Cook's Synthesis Toolkit in C++ [8].

This voice synthesizer, in order to produce a "petasti"-like phonation quality, produced discrete sounds like the ones shown in Fig. 7(a). If we consider that phoneme /o/ has the typical formant distribution described in Table I, then the conventional synthetic reproduction of /o/ with "petasti" would merely yield three intermediate notes, with the second note raised by a semitone. This means that a quarter-note influenced by "petasti" is analyzed as the sequence [sixteenth-note, eight-note sharp, sixteenth-note.]

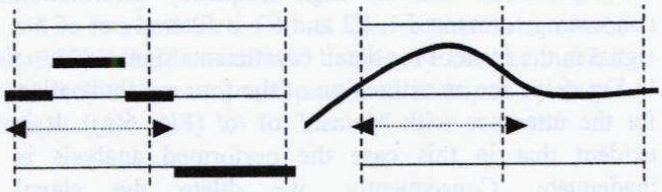

Fig. 7. (a) Quantum leap versus (b) continuous F0 synthetic performance. The arrow heads denote the endpoints of the transient phenomenon. The horizontal grid lines denote whole tone intervals while the vertical grid lines set the time units in quarter-notes (♪).

The improvement we have made in order to reproduce "petasti" more naturally relies on the shape of the F0 curve that we have identified in the previous section. Accordingly, we simulate "petasti" with a Gaussian having its peak raised by a semitone. The time duration of the Gaussian, estimated by the width of curve at –3dB below its peak point, is slightly bigger than a time unit. The F0 curve that is produced for an interval of a major second (i.e. a whole tone) is shown in Fig. 7(b).

Of course, the synthesizer cannot produce an utterance with a continuous F0 fluctuation. However, it can invoke very short bursts of impulses, lasting 10 ms each, which can approximate adequately the continuous curve.

Apart from the F0 fluctuation the F1, F2 and F3 are also affected. Although their influence is not significant, yet they add naturalness to the synthesized phonemes.

Phoneme	formants (Hz)	Freq. Width	Relative amplitudes (dB)
/o/	515	0.977	0
	1805	0.810	-10
	2526	0.875	-10

Table I. Estimated mean values for phoneme /o/. The frequency width of the formants is not denoted in Hz but as it is inserted into the synthesizer, i.e. exp(2W/Fp).

REFERENCES

[1] "Monumenta Musicae Byzantinae", Institute for Greek and Latin, University of Copenhagen. http://www.igl.ku.dk/MMB.

[2] Manuscript from Iveron Monastery of Mount Athos, cod. 437, 1805 AD, reproducing the Doxastarion of Jacov protopsalt (†1800).

[3] Panagiotopoulos, D., *Theory and Praxis of Byzantine Ecclesiastical Music*, 3rd ed, Athens, 1982 (In Greek).

[4] Jacobs, A., *The Penguin Dictionary of Music*, Penguin, USA, 1991.

[5] Pikrakis, A., Theodoridis, S., Kamarotos, D., "Recognition of Isolated Musical Patterns in the context of Greek Traditional Music", *Third IEEE International Conference on Electronics, Circuits and Systems ICECS '96*, Rhodes, pp. 1223-1226, October 13-16, 1996.

[6] Politis, D., Tsoukalas, A., Linardis, P., Bakalakos, A., "VIDI-A Voice Instrument Digital Interface for Byzantine Music", *International Computer Music Conference ICMC97*, Thessaloniki, Proceedings, pp. 403-407, September 25-30, 1997.

[7] Strang, G., Nguyen, T., *Wavelets and Filter Banks*, Wellesley-Cambridge Press, 1996.

[8] Cook, P., "Synthesis Toolkit in C++", Department of Computer Science, Princeton University, USA. http://www.cs.princeton.edu/~prc

АВТОМАТИЗАЦІЯ КРИМІНАЛІСТИЧНИХ ДОСЛІДЖЕНЬ МОВЛЕННЄВИХ СИГНАЛІВ

В.М.Магера І.І.Горбань, С.В.Лєвий

Державний науково-дослідний експертно-криміналістичний центр МВС України, Україна, 01024, м. Київ, вул. Богомольця, 10, тел./факс.: 291-39-53.

Іститут проблем математичних машин і систем Національної Академії наук України, Україна, 01187, м. Київ, проспект Глушкова, 42, електронна пошта: gorban@immsp.kiev.ua, Тел.: 38(044) 266-61-74, факс:.38(044) 446-8129.

Галузева науково-дослідна лабораторія методів і засобів спеціального призначення радіотехнічного факультету Національного технічного університету України, кафедра радіотехнічних приладів і систем, Україна, 01056, м. Київ, пр-т Перемоги, 37 (2103), електронна пошта: Leviy@lmmfr.kiev.ua, тел./факс.: 274-89-84.

АНОТАЦІЯ

Описано програмно-апаратний комплекс та програму, що дозволяють досліджувати мовленнєві сигнали у автоматичних та автоматизованих режимах. Висока ефективність досліджень досягається на сигналах із рівнем спотворення до 35 дБ та відношенням сигнал-шум – до 10-12 ДБ. Наводиться методика і результати тестування програм.

ВСТУП

Сучасний стан розвитку апаратури запису мовлення вимагає адекватного розвитку засобів і методик криміналістичних досліджень. Програмно-апаратні засоби, які зараз використовуються для криміналістичних досліджень, вже не задовольняють сучасним вимогам. Наприклад, програми для запису та редагування звукових сигналів "Cool Edit Pro" фірми "Syntrillium Software Corporetion" США [1] і "Sounde Forge 4.5" фірми "Sonic Foundry, Inc." США [2] дозволяють, головним чином, здійснювати редагування звукових сигналів та знижувати рівень шумів. Однак дослідження характеристик звукових сигналів в них представлено дуже слабо. Апаратно-програмні засоби "Computerized Speech Lab" фірми "Kay Elemetrics Corp." США [3] і "SIS" – ТОВ "Центр речевых технологий" м. Санкт-Петербург [4] дозволяють всебічно досліджувати різні характеристики мовленнєвих сигналів, проводити статистичний аналіз, але вони мають низький рівень автоматизації проведення досліджень. Програма "Диалект" м. Москва [5] дозволяє в автоматизованому режимі проводити ідентифікаційний аналіз мовленнєвих сигналів. Але великий обсяг підготовчих операцій і оцінка деяких проміжних результатів обчислень покладається на оператора.

Зазначені програмні засоби та деякі інші демонструють гарні результати лише при роботі в умовах малих частотних спотворень і низького рівня перешкод. Коли деструктивні фактори значні, ці системи не ефективні.

Намагання створити нове покоління систем дослідження мовлення істотно більш ефективних, ніж існуючі, привело до появи нових ідей, методів і алгоритмів обробки, орієнтованих на підвищення стійкості роботи в умовах спотворень і зашумлення мовленнєвих сигналів [6-12]. Аналіз і тестування декотрих нових алгоритмів розпізнавання особи по мовленню виявили загальний їх недолік: алгоритми достатньо ефективні, коли частотні спотворення не перевищують 10-15 дБ і спектр реальної завади співпадає з очікуваним. У наведеному нижче матеріалі описані дві системи, що в значній мірі позбавлені вказаних недоліків. Це автоматизований програмно-апаратний комплекс "Phonograph" та автоматична система CASVI (Crime-detection Automatic Speaker Verification and Identification System).

1. Автоматизований програмно-апаратний комплекс "Phonograph"

Автоматизований програмно-апаратний комплекс "Phonograph" призначений для криміналістичних досліджень в мовленнєвому діапазоні частот: голосу, сигналів акустичної і неакустичної природи виникнення, шумів і завад. Він входить до складу автоматизованої криміналістичної системи "Логос" для дослідження матеріалів та засобів звуко- і відеозапису [13].

Комплекс складається із персонального комп`ютера типу Pentium II, звукової карти ExpertColor MED 3201 InterWave AMD

/AM78C210KC/, узгоджувального пристрою "Sound card satellite" (SCS), комутатора зовнішніх пристроїв КВУ-54, принтера, головних телефонів "AKG K141", магнітофонів (котушкового "Электроніка-004К-Стерео", мікрокасетного "Panasonic RR-930" і компакт-касетного "TEAC V-2030S"), відеомагнітофона "Panasonic NV-F55", підсилювача "Kenwood", акустичної системи "Амфітон 100АС-022" і програмного забезпечення.

Структурна схема автоматизованого програмно-апаратного комплексу "Phonograph" наведена на рис.1.

Відтворення і запис звукових файлів в(із) пам`яті комп`ютера проводиться з використанням спеціально розробленого пристрою SCS, який узгоджує по рівню вихідний канал магнітофону і лінійний вхід звукової плати. Пристрій дозволяє прослуховувати лівий, правий стереоканали і здійснює їх змішування. Конструктивно він розміщується в великому (5,25") зовнішньому відсіку, який виходить на лицьову сторону персонального комп`ютеру, що забезпечує вільний доступ до всіх органів регулювання.

Комутатор зовнішніх пристроїв здійснює: комутацію 16 звукових лінійних стерео входів з 16 звуковими стерео виходами; комутацію джерела звукового сигналу для прослуховування на головні телефони через вбудований стерео підсилювач низьких частот в дихотонічному режимі (комутація лівого телефону навушника незалежна від комутації правого); перетворення стерео виходів в моно виходи; комутацію одного із 4-х відео входів з одним відео виходом; включення/виключення мереженої напруги 220 В, 50 Гц 16-ти зовнішнім пристроям. Управління всіма режимами роботи комутатору проводиться від персонального комп`ютеру. Створення комутатору дозволила значно підвищити рівень автоматизації і зв`язати всі периферійні пристрої в функціонально єдиний комплекс.

Програмне забезпечення "Phonograph" дозволяє здійснювати: ввід-вивід аналогового сигналу в(із) ПЕОМ; функції цифрового магнітофону; фільтрацію і компандування фонограм в реальному часі; виміри часових, амплітудних, спектральних і кепстральних характеристик мовленнєвих і інших сигналів; виконання статистичного аналізу характеристик сигналів; автоматизований ідентифікаційний аналіз і отримання оцінки співставлення голосів різних дикторів; встановлення номеру копії фонограми по частоті мережі 50 Гц; автоматичне встановлення телефонного номеру по тональним посиланням АТС, які зафіксовані на фонограмі та інше.

Комплекс забезпечує виконання досліджень у відповідності до створених та затверджених методик.

Ідентифікаційні дослідження мовлення представлені методикою аналізу акустичних характеристик голосу і лінгвістичною методикою аналізу мовлення. Дослідження акустичних характеристик голосу виконується за традиційним (дослідження окремих часових, амплітудних, спектральних, кепстральних та інших характеристик мовленнєвого сигналу) і автоматизованим аналізами. Кількість ознак, що досліджуються, при традиційному аналізі становить 27, при автоматизованому – більше 100. Автоматизований аналіз виконується на довільному тексті та на текстозалежному матеріалі. Лінгвістичний аналіз виконується за просодичними та фонетичними характеристиками мовлення. Кількість ознак, що досліджуються, при лінгвістичному аналізі - 32.

У комплексі "Phonograph" автоматизовано технологічні операції визначення характеристик мовлення на придатність для проведення ідентифікаційних досліджень, відбір для аналізу подібних голосних звуків, розмітку мовленнєвого сигналу, розрахунок характеристик по окремим ознакам мовленнєвого сигналу, формування рішення і інше.

Комплекс розрахований на кваліфікованих операторів, які працюють у сфері дослідження матеріалів та засобів звуко- і відеозапису.

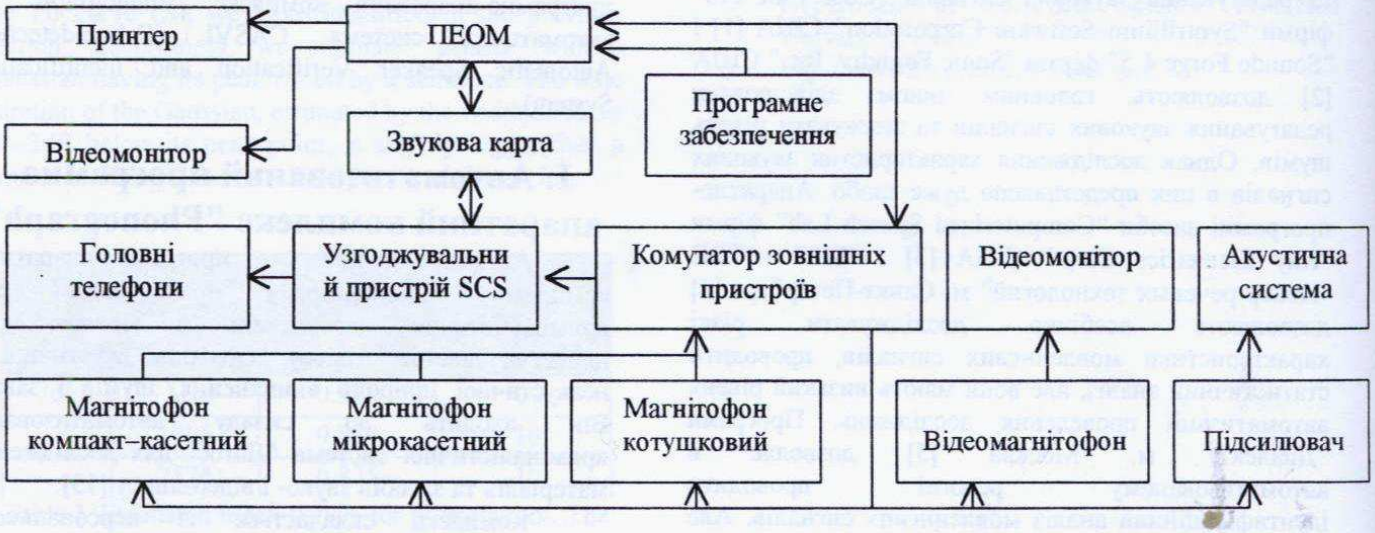

Рис.1. Структурна схема автоматизованого програмно-апаратного комплексу "Phonograph"

2. Система CASVI

Система CASVI забезпечує ідентифікацію і верифікацію особи за голосом у автоматичному режимі. Загальний опис цієї програми і її можливостей наведено в роботах [14-17]. Система ефективного вирішує задачі, що на неї покладаються, при наявності великих частотних спотворень і зашумленні мовленнєвих сигналів. Головним її елементом є підсистема ідентифікації (рис. 2). Вона складається із системи функціональних програм (СФП), системи управління (СУ) і бази даних (БД).

СФП забезпечує обробку і порівняння сигналів і включає програму розрахунку спектрів (РС), програму розрахунку інформаційних ознак (РІО) і програму порівняння інформаційних ознак (ПІО).

СУ виконує задачі взаємодії з користувачем, управління процесами обробки і порівняння фонограм, формування і поповнення бази даних. Вона містить програму, яка забезпечує інтерфейс для користувача (ІК), програму управління функціональними програмами (УФП) і програму управління базою даних (УБД).

Час, необхідний для прийняття рішення підпрограмою, визначається режимами роботи і можливостями комп`ютеру і може становити від декілька хвилин в простих випадках до десятків хвилин.

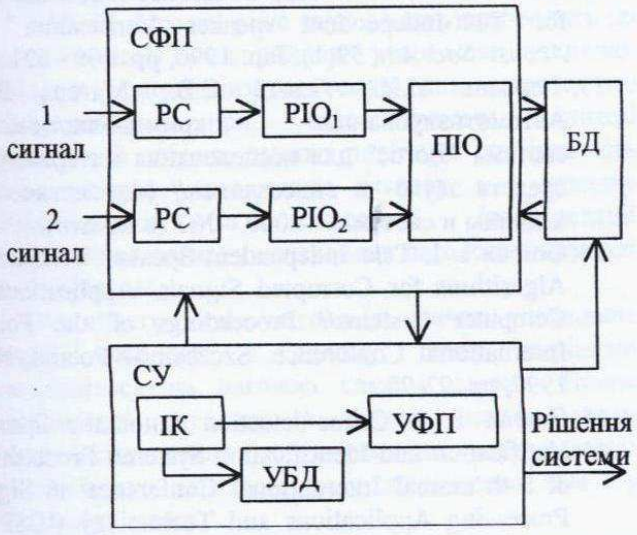

Рис. 2. Структура підпрограми ідентифікації CASVI

Функціональні програми реалізують кепстральні алгоритми.

РС проводить розділення запису на окремі короткі фрагменти (фрейми), для кожного фрейму розраховує спектр, статистичну характеристику корисного сигналу і завади та сортує фрейми на сигнальні і завадові. РІО реалізує формування сигнальних блоків фреймів і розрахунок інформаційних ознак: центрування кепстрів, а також перші і другі похідні кепстрів.

ПІО проводить розрахунок відстані між записами по кожній інформаційній ознаці окремо, розрахунок оцінок щільностей імовірності для однакових і різних голосів із БД, формування підсумкової відстані між голосами, формування порогу з використанням отриманих оцінок щільностей імовірності, порівняння підсумкової відстані з порогом, формування рішення про ідентичність голосів і розрахунок оцінки імовірності помилкового рішення.

В СУ головними є програми, які забезпечують ІК і УФП. Інтерфейс користувача СУ складається із вікон головного меню, поточного результату та бази даних.

Головне меню дозволяє вибрати файли, оцінити рівень шумів фонограм, прослухати їх і виконати ідентифікацію по одному із описаних вище варіантів.

Вікно поточного результату надає результат ідентифікації у вигляді рішення системи Yes/No (да/нет) і розрахункову імовірність помилки прийнятого рішення. Воно інформує користувача про рівень шумів фонограм.

Вікно бази даних дозволяє відібрати файли для БД, прослухати їх і запустити розрахунок *спектрів і кепстральних ознак, які використовуються* при ідентифікації.

Підсистема ідентифікації CASVI розрахована на роботу з тексто-незалежним матеріалом. Але, враховуючи низьку якість запису реальних фонограм, які надходять на експертизу, в окремих випадках доцільно використовувати для порівняння заздалегідь підготовлений матеріал. У цьому випадку з допомогою комплексу "Phonograph" у автоматизованому режимі може проводитись відбір подібних наголошених голосних, що стоять між приголосними звуками. Із них формуються масиви для порівняння з допомогою підсистеми ідентифікації CASVI.

Програма орієнтована на операторів з нижчим ступенем кваліфікації, ніж для роботи з комплексом "Phonograph", так як більшість технологічних операцій повністю автоматизована. Це зменшує витрати на спеціальну підготовку експертів.

3. Тестування працездатності комплексу "Phonograph" та системи CASVI

Тестування програмно-апаратного комплексу "Phonograph" здійснювалось згідно з програмою, затвердженою Секцією експертизи матеріалів та засобів звуко- і відеозапису Науково-методичної ради Міністерства юстиції України. Для кожного окремого режиму чи розрахунку були сформовані спеціальні сигнали та вхідні дані, на яких перевірялась працездатність комплексу. Тестування показало відповідність усіх параметрі розрахунковим і дієздатність всієї системи в цілому. Комплекс "Phonograph" був розглянутий на Науково-координаційній раді МВС України та рекомендований до впровадження у експертну практику.

Алгоритми ідентифікації CASVI тестувались на 32 п`ятихвилинних записах чоловічих голосів. Відношення сигнал-завада для всіх записів складало

не менше 30 дБ. Записи були розділені на дві однакові групи, до яких входили по два комплекти записів восьми дикторів. Записи першої групи використовувались як зразки, а другої - як робочі. Ефективність роботи досліджувалась при сильних частотних викривленнях сигналу і зашумленні, типових при передачі повідомлення по телефонному каналу. Для цього перший комплект робочих записів викривлявся по частоті і зашумлявся на кожній частоті до фіксованого рівня.

Експерименти проводились в два етапи: навчання і тестування. На етапі навчання формувались дані для БД з використанням зразкових записів. Ефективність алгоритмів перевірялась на етапі тестування з використанням робочих записів. Число спроб в разі однакових дикторів було 72, у випадку різних - 504. Тривалість записів дорівнювалась 12 секундам.

Результати експериментів показали, що в умовах частотних викривлень до 35 дБ і зашумленні до відношення сигнал-завада 12 дБ імовірність правильної ідентифікації становить 90% при імовірності помилкової тривоги 10%.

Висновки

Результати тестування програмно-апаратного комплексу "Phonograph" та системи CASVI свідчать про їх ефективну роботу при дослідженні мовленнєвих сигналів.

Програмно-апаратний комплекс "Phonograph" є більш дослідницькою програмою і дозволяє виконувати всі види акустичних та просодичні і фонетичні лінгвістичні дослідження. Значна частина технологічних операцій та режимів автоматизована, що значно скорочує час виконання експертиз.

Практично всі функції підпрограми ідентифікації CASVI працюють у автоматичному режимі при значних частотних викривленнях (до 35 дБ) і відношенню сигнал-шум 12 дБ. Наведені результати суттєво перевищують ті, що до цього часу було досягнуто. Значне покращання технічних параметрів стало можливим завдяки комплексному підходу до проблеми і використанню нових робастних алгоритмів обробки, оптимізованих для роботи в несприятливих умовах.

Комплекс "Phonograph" та система CASVI за результатами порівняння записів голосів наводять не тільки рішення про їх ідентичність, але і оцінку імовірності помилки прийнятого рішення.

ЛІТЕРАТУРА

1. http://www.syntrillium.com
2. http://www.sfoundry.com/index.html
3. Новосельский А.Ф. Компьютерная система для всесторонней работы с речью CLS MODEL 4300B// Компьютеры+Программы. – 1995. – №7(22). – С.66-71.
4. Рекомендации по эффективному использованию возможностей АРМЭФ SIS при выполнении криминалистических экспертиз. – Центр речевых технологий. – Санкт-Петербург, 1995.
5. Идентификация лиц по фонограммам русской речи на автоматизированной системе "Диалект" / Попов Н.Ф., Линьков А.Н., Кураченкова Н.В., Байчаров Н.В. / Под ред. Фесенко А.В. – М.: В/ч 34435, 1996. – 102 с.
6. Hansen J. H. L, Mammone R. J., Young S. editors// IEEE Transactions on Speech and audio Processing. - 1994. – October.
7. Mansour D. and Juang B. H. The short-time modified coherence representation and noisy speech recognition// IEEE Trans. Acoust., Speech, Signal Processing. –1989. –37. – June. – P. 796-804.
8. Newneyer L. and Weintraub M. Probabilistic Optimum Filtering for Robust Speech Recognition// Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing. – 1994. –1. – P. 417-420.
9. Reynolds D. A. and Rose R. C. Robust Text-independent Speaker Identification Using Gaussian Mixture Speaker Models// IEEE Trans. Speech, Audio Processing. –1995. – 3. – January. – P. 72-83.
10. Nolazco Flores J. A., Young S. J. Continuous Speech Recognition in Noise Using Spectral Subtraction and HMM Adaptation// Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing. – 1994. –1. – P. 409-412.
11. J. Mammone, X. Zhang, R. P. Ramachandran, "Robust Speaker Recognition," *IEEE Signal Processing Magazine,* Sept. 1996, pp. 58 - 71.
12. M. A. Lund, C. C. Lee, "A Robust Sequential Test for Text-Independent speaker Verification," *J. Acoust. Soc. Am.* 99(1), Jan. 1996, pp. 609 - 621.
13. Горбань И.И., Левый С.В., Магера В.Н. Автоматизированная криминалистическая система "Логос" для исследования материалов и средств звуко- и видеозаписи// Математические машины и системы. –2000. –№3 (в печати).
14. Gorban I. I. Text Independent Speaker Verification Algorithms for Corrupted Signals. Applications of Computer Systems// Proceedings of the Fourth International Conference. Szczecin – Poland, Nov, 1997, pp. 92-98.
15. Gorban 1. 1. Crime-detection Automatic Speaker Verification and Identification System// Proceedings of 9-th annual International Conference on Signal Processing Applications and Technology (ICSPAT 98). - 1998. - August.
16. Горбань И.И., Горбань Н.И., Клименко А.В., Хазанович М.С. Подсистема верификации новой криминалистической системы автоматической верификации и идентификации личности по голосу (CASVI)// Математические машины и системы. – 1997. –№2. –С.61-64.
17. Горбань И.И., Клименко А.В. Робастні алгоритми верифікації особи за голосом, що призначені для роботи в умовах сильних завад та спотворень мовних повідомлень// 4 Всеукраїнська міжнародна конференція по обробці та розпізнаванню образів "Укробраз 98". – 1998.

ГЕНЕРУВАННЯ ПРАВИЛ РОЗСТАВЛЯННЯ НАГОЛОСІВ У БАГАТОМОВНОМУ АСПЕКТІ

Микола Сажок

Міжнародний науково-навчальний центр інформаційних технологій та систем
40 просп. Академіка Глушкова, Київ 252022
Електронна пошта: mykola@uasoiro.freenet.kiev.ua

Mykola Sazhok. **Generating Stress Rules in Multilingual Aspect.** The risen problem of stresses, particularly in Ukrainian, often confuses text-to-speech and recognition engines. Given analysis of stress in Ukrainian generalizes knowledge about this important aspect. Cases when stress moving in a word changes its meaning and/or its role in sentence are considered. Revealed properties allow for building heuristics and statistics based rules predicting a stress position in most cases and resolving several types of ambigual stresses in words. Within entire paper cross-lingual stress aspects are discussed on example of Ukrainian and English. It is shown how parts of word influent on stress position and how narrow context helps to resolve stress ambiguities.

1 Вступ

Роль наголосів в усній мові важко переоцінити. Властивість наголосу коригувати та й зовсім змінювати зміст промовленого широко використовується у спілкуванні людей.

Хибний наголос ріже слух і часто призводить до непорозуміння при сприйнятті усної мови людиною. Тому в усномовних системах для практичного використання при автоматичному синтезі правильний наголос значно поліпшує розбірливість синтезованого усномовного сигналу, а у випадку розпізнавання та смислової інтерпретації усномовного сигналу нехтування наголосами взагалі неприпустиме.

Розмаїтість типів наголосів значно ускладнює аналіз наголосу в цілому, тому вирішено зосередитися на наголосі саме у словах. Попри особливості наголосу в різних мовах, можна виокремити чимало спільних рис, притаманних розставленню наголосів як у споріднених, так і у віддалених мовах.

Використання знань про наголос в системах автоматичного синтезу, розпізнавання та інтерпретації усномовного сигналу завжди було на часі. Проте, для мов, де наголос є надто нерегулярним і не позначається на письмі, зокрема для української та англійської, дослідження з цього питання потребують більше передумов та ретельності. Особливо це стосується для мов, що допускають велику кількість словоформ, до яких відноситься українська.

2 Загальний аналіз природи наголосу

Що ж таке по суті є наголос? Чіткого визначення цієї просодичної характеристики напевно не існує. Очевидно, що це явище є чимось принципово важливим в усній мові. Також помічено, що під час наголошення вимовляння найбільш відповідає фонетичним нормам та є найближчим до написання. Тобто в наголошеній позиції фонема-алофон звучить (реалізується) найбільш чітко, найменшою мірою піддаючись спотворенням та редукції.

В загальному випадку, наголос може бути сильнішим та слабшим, виокремлюючись на тлі як цілої фрази, речення, так і лише окремого слова або сусідніх складів. Отже, наголоси формуються на трьох рівнях: фрази, слова або синтагми та на субсловарному мікрорівні або на рівні складів.

Зосереджуючись на рівні не вище слова, зазначимо, що тут наголоси бувають основні та другорядні, які балансують на межі мікрорівня.

Позиція основного наголосу у слові рахується мовознавцями в залежності від мови частіше з кінця слова, але буває і з початку. Тут у деяких мов спостерігається регулярність. Хоча при вимовлянні фрази в цілому слово може втратити наголос, набуваючи другорядного наголосу зовсім не там, де розташовується основний. Типовою ілюстрацією такої властивості є приклад з французької "m*ou*len r**ou**ge", де жирні літери позначають основний наголос, а похилі літери – другорядний.

Простежується також залежність наголосу від типу мовлення. В першу чергу ритм, потім темп, розмовність впливають на формування наголосів, їх плавання, перехід у слабкіший стан. Вплив ритму на наголоси чудово простежується на зразках поезії, пісенної творчості, коли диктат ритму інколи зовсім не зважає на граматично правильний наголос.

Прикладом цього є уривок з української народної пісні: "Ой у лісі калина, калина, калина", у той час, як словник подає "калина"

В українській мові основний наголос може перебувати на останньому або другому з кінця складі. Досить поширений також дактилічний наголос, тобто третій з кінця. Але цим не обмежується і в деяких словоформах позиція наголосу може перебувати аж на сьомому з кінця складі як-от у слові "спізнюватимемося".

Англійська мова також має нерегулярний наголос. Відлік позиції наголосу ведеться з початку слів, як для більшості германських мов, і припадає основний наголос переважно на перший склад. Помічено також тенденцію англійською мови до перероблення наголосів запозичених слів, наприклад французьких, на свій лад.

Наголос тісно пов'язаний з іншими просодичними характеристиками, особливо інтона-

ційними. Простежується чітка залежність різких змін інтонаційного контуру саме одночасно з початком наголошеного складу. Безумовно, кожна усна мова своєрідно реалізує наголос, проте крос-лінгвістичні порівняння в багатьох випадках дозволяють адаптувати знання про наголос з однієї мову на іншу.

3 Автоматичне розставлення наголосів

Побудувати алгоритм розставлення наголосів для мов з нерегулярним наголосом у словах і просто і складно. Здавалося б, що може бути простішого, ніж знайти слово з тексту в словнику, де завчасно проставлені наголоси та поставити наголос у шуканому слові. Проте такий шлях вирішення не може задовольнити з багатьох причин.

Для мов з розмаїтістю словоформ, якою є українська, де наголоси від форми до форми часто змінюються, саме створення словників, які враховують позицію наголосу є проблематичним, не кажучи вже про обсяги пам'яті, потрібні для такого словника, та час доступу до певного слова.

Потреба у використанні величезних обсягів пам'яті обмежує використання усномовних технологій як на сучасних комп'ютерах, так і в портативних виробах особливо, коли обмеження на обсяги пам'яті та потужності процесора суттєві. Вихід з цього закуту вбачається у винайденні правил, які б дозволяли розставляти наголоси в довільному слові для певної мови.

Спроби формування таких правил вперше були зроблені для англійської мови. Як було зазначено, велика кількість англійських слів містить основний наголос на першому складі. Водночас були виявлені суфікси, що "перетягають" наголос. Це "-ion", "-ize", "-ate" та деякі інші. Послабленими в багатьох випадках залишаються також префікси. Так на прикладі слів "expect" і "expectation" видно, що префікс "-ex" залишається ненаголошеним, а додавання суфіксів "-ate", "-ion" веде до зсуву наголосу в напрямку до суфікса.

В українській мові на роль найбільш поширеного наголосу претендують як перший, так і другий і третій від останнього. Остаточно визначити це можливо лише проаналізувавши ужиток слів, керуючись текстовим корпусом, а також з міркувань більш економного задання правил. Також виокремлюються нечисленні суфікси, за якими можна визначити наголос ("-еньк", "-есенък"). Крім того, спостерігається чітка тенденція до послаблення закінчень, за винятком прикметників.

Складені та довгі слова видаються найбільш проблематичними для розставлення наголосу, особливо в англійській мові.

4 Зсув наголосу при творенні слів і форм

Зміна наголосу широко використовується при словотворенні. Так англійські "record" і "record" ілюструють перехід іменника у дієслово, під час якого відбувається зсув наголосу.

Українська мова більш широко використовує зміни позиції наголосу і не лише при творенні нових слів, але і при творенні форм слова. Так в іменниках наголос може зсунутись при переході у множину (рука-руки, поле-поля), рідше при звичайному відмінюванні під впливом певних суфіксів.

Дієслова досить вільно поводяться з наголосами, хоча простежуються деякі закономірності з урахуванням сукупного впливу на наголос усіх частин слова та кількості складів у слові.

Прикметники найбільш консервативно зберігають наголошену голосну, при тому, що утворені від них прислівники уникають наголосів на останньому складі (значний-значно), інколи – на передостанньому (веселий-весело).

5 Проблема неоднозначності наголосів

Зміна наголосу в слові призводить, як було сказано, найчастіше до помилки у вимові. Випадки, коли зміна наголосу веде до переведення слова в іншу словоформу, або змінює його значення на інше, часом безпосередньо не пов'язане з початковим семантично, призводять до помилки у смисловій інтерпретації як окремого слова, так і цілого вислову.

У випадках, коли наголос змінює форму слова слід витримати непорушність узгодження слів у реченні. Для цього достатньо визначити керування в якому перебуває це слово за допомогою слів-орієнтирів з найближчого контексту. Такими орієнтирами є прийменники або їх відсутність, прикметники, як найбільш консервативні у наголошенні, та слова з інших частин мови, які вимагають певного узгодження, і чиє значення не викликає сумніву.

Так, у словосполученні "знімати з руки" прийменник "з" вимагає після себе родовий або твірний відмінок, тому й наголос має відповідати одному з цих відмінків. У вислові "потисли один одному руки" слово "руки" керується дієсловом "потисли", яке вимагає знахідного відмінку.

У випадках, коли можливі варіянти наголосів, що відповідають різним частинам мови, постає проблема визначення частини мови цього слова в контексті. Варіянти наголосів, які не змінюють частину мови, змінюючи при цьому значення слова (плачу, замок), слід розглядати окремо.

6 Підсумок

Проведений аналіз дозволяє виробити евристичні та статистичні підходи до формування правил розставлення наголосів не лише в контексті слова, а і словосполучення та цілої фрази, розв'язуючи у багатьох випадках проблему неоднозначності інтерпретації слів при усномовному синтезі та розпізнаванні. Для багатомовних систем важливим напрямком подальших досліджень вбачається використання знань про наголоси зі споріднених мов.