IRTC Activity in Speech Information Technology towards East-West Europe Co-Operation

Taras K. Vintsiuk

International Research-Training Centre for Information Technologies and Systems – IRTC
40 prospekt Akademika Hlushkova, Kyiv 03680, Ukraine
vintsiuk@uasoiro.org.ua

Abstract

Research and development projects in speech technologies, which carried out since 1986 at the NAS Institute of Cybernetics (IC) and then from 1997 continued at IRTC, are analysed. It is considered the efforts coordinated and/or promoted with UNESCO, ELSNET and INTAS. Then INTAS project titled “Language Independent Model for ASR Advanced Research” is declared. Examples of created portable devices for IST are given.

1. Speech dialogue system “RECH-121”

In 1986–1989 accordingly with the contracts between UNESCO and Institute of Cybernetics it was been created the Multilingual Speech Dialogue System (MSDS) of RECH series, model 121 (Figure 1) [1].

This system dealt with 7 languages (Ukrainian, Russian, English, French, Spanish, German, Italian). The volume of an operating vocabulary was not more than 256x3=768 words. MSDS recognised and translated (word-by-word) the words pronounced separately and continue speech, which is composed from the words of the chosen vocabulary. The permissible phrase duration was equal 20 s. The recognition response time delay did not depend on phrase duration and was equal 0.3 s.

As for speech synthesis the best results were obtained for Ukrainian and Russian. The phonemes of these languages were used as the base in speech synthesis for other languages. While speech was being synthesised, a cartoon face moved synchronously with “pronounced” phones.

MSDS consisted of two blocks: PC 286 and proper SDS as speech input-output device controlled by PC.

MSDS “RECH-121” functioned in complex with CDS/ISIS micro, developed and distributed by UNESCO.

2. INTAS project 93–2119

During 1993–1996 we took part in the INTAS project 93–2119 “Extension of ELSNET to the NIS”. The University of Edinburgh, Human Communication Research Centre, Language Technology Group was the Coordinator. Another INTAS country partner was Centre National de la Recherche Scientifique, Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur – LIMSI.

Through this project ELSNET has incorporated two new nodes in the NIS (NAS Institute of Cybernetics, Speech Science and Technology Department, Kyiv, Ukraine and RAN Institute for Information Problems, Creative Research Laboratory, Moscow, Russia) to the European Network in Language and Speech (ELSNET). Because these NIS research groups “have excellent research track records in the NL processing and speech areas. They have been active in international communication and co-operation but their links with the research community in Western Europe have been sporadic and incidental” [2].

Owing to the project our research group had gained considerable scientific leverage by exchanging information, research results, data and tools with the rest of the ELSNET community.

Apart from the development of tools, this collaboration has provided an active framework for the integration of the research group in the Ukraine into the European scientific community.

NAS IC surveyed all language and speech groups in the Ukraine. It logged information about the theoretical platforms used by the researchers in their work, about computational tools developed locally, about publicly available tools and resources used by the groups, etc. This information could be used to update ELSNET’s current report on "Profiles of Language Engineering Organisations in Central and Eastern Europe and Selected New Independent States" (August 1994). The conclusions of the survey also gave a more general overview of the state of NLP research in Russia and Ukraine and were used as input to various of ELSNET's strategic initiatives, in particular concerning the development of a Pan-European Network in Language and Speech.

A second goal of the project was to carry out collaborative research and development work, which since 1997 is continued at IRTC that is ELSNET academic participant.

Results of the collaborative work and its extension have been reported and presented in various articles and conferences organised and/or sponsored by INTAS and ELSNET:

- “ELSNET Goes East and EMACS Workshop”, Moscow 1996;
- SpeCom’97, Cluj-Napoca, Romania 1997;
- SpeCom’98, St-Petersburg 1998;
- SpeCom’99, Moscow 1999 [3–7].

Now let us summarise the obtained theoretical results, from which the new INTAS proposal is derived.

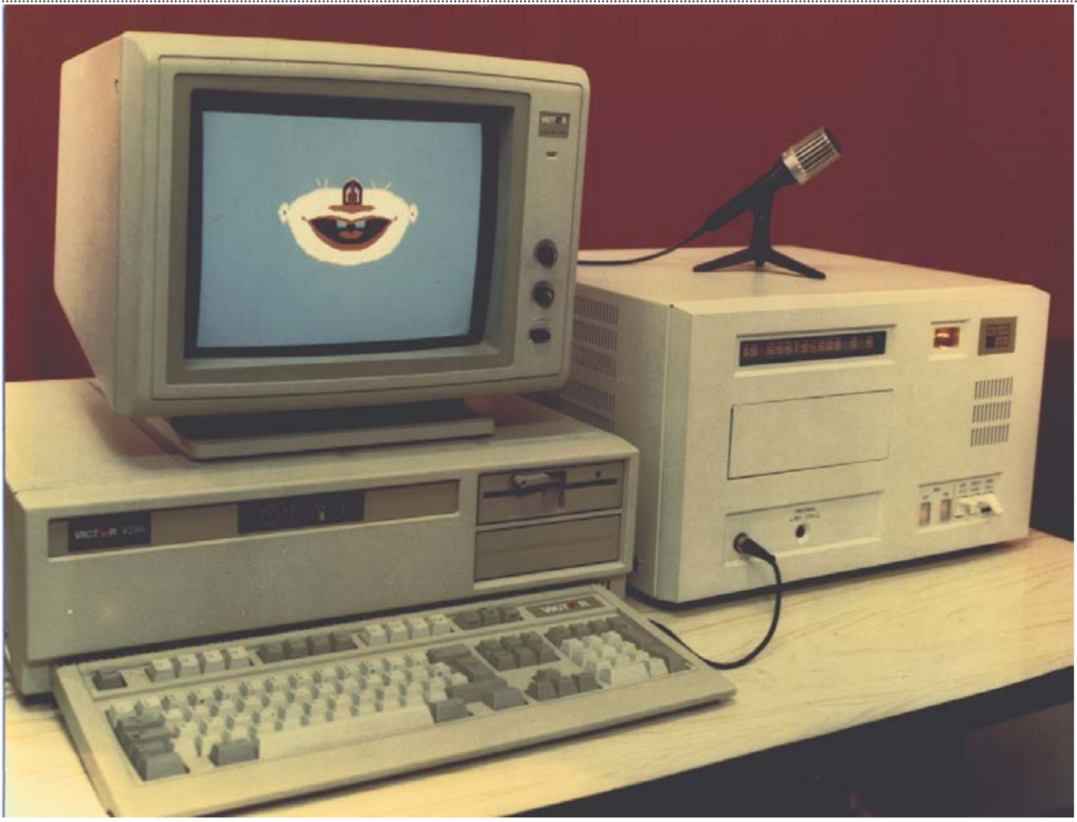

Figure 1. Multilingual speech dialogue system RECH-121 (1986).

3. Generative model

In [3–4] two models of a Dictation/Translation Machine (DTM) have been proposed. They are hierarchically organised. The first approach is based on so-called generative model of speech signal recognition, understanding and synthesis. The DTM generalised structure based on this model is shown in Figure 2. Here a speech synthesis is used as a feedback in the speech recognition process.

The three main parts of the DTM are 1) external world model which describes all possible meanings to be transmitted in communication, 2) natural language texts/sentences generator, and 3) phonetic-acoustic speech signal prototypes generator. The problem of automatic spoken DTM is formulated as 1) a problem of *finding*, for the recognisable signal, *the most similar connected-speech signal prototype* from the set of all possible prototype signals generated by the third part for all possible natural language texts and sentences, specified by the second part for all possible meanings to be transmitted which are determined by the first block, and as 2) a problem of *analysing (examining) the latter*, as a series of words and a canonical form of meaning carried by speech signal. The delivered series of words are grammatically and semantically valid and therefore can be printed (dictation machine) or synthesised by the three-part channel for other language (translation machine).

The DTM main part is the block (model) of an External World that is a EW-model. For instance the EW-model can be considered as a union of sub-models for different subject areas, with the using of a common part, which expresses general qualities of the external world. The EW-model is joint for all natural and artificial languages and as a matter of fact depends little of them. The EW-model is specified by a particular mathematical language e.g. by a canonical forms language. Falling away details here let us emphasize only that the EW-model generates canonical forms of the meanings to be transmitted in speech communication process. Evidently for any subject area it is possible to determine a finite set of canonical forms, which are feasible in dialogue. For example when it is saying about the spoken calculator with only the four rules of arithmetic then in this case the permissible canonical forms have the simplest structure: these are operation and first and second operands.

The next, second in the hierarchy, part of the DTM is a generator of Natural Language Texts or Sentences that is NLTS-generator. This block is based on the knowledge of vocabulary, word-building, morphology, phraseology and syntax for the concrete natural language. Receiving a canonical form of the meaning to be transmitted from the external world model the NLTS-generator synthesises all permissible for this language texts and sentences, which express the same meaning designated by the accepted canonical form. It is possible to say that the NLTS-generator is a semantic-grammatical model of the natural language.

At the third level of the hierarchy the Phonetics-Acoustic Speech Model Signal generator is found that is

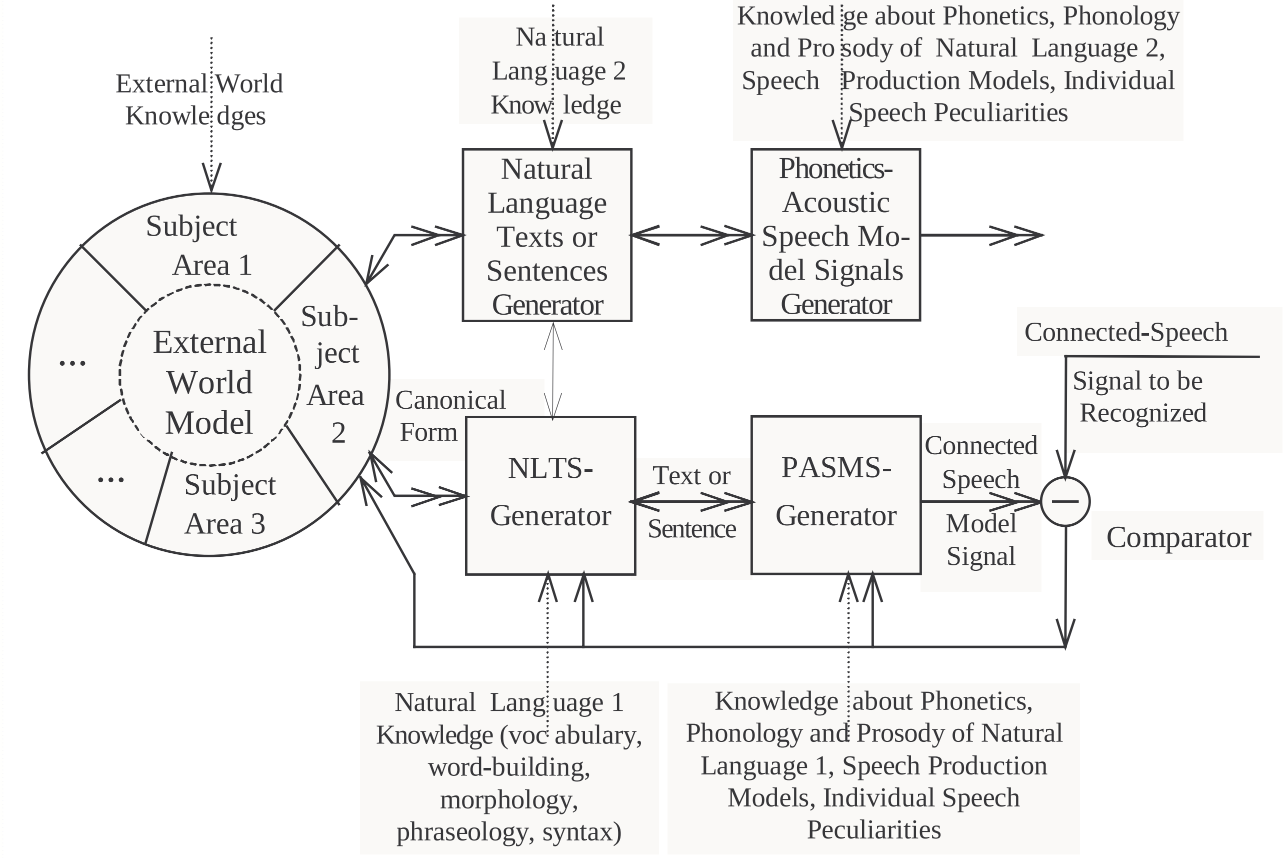
External World Knowledges
Natural Language 2 Knowledge
Knowledge about Phonetics, Phonology and Prosody of Natural Language 2, Speech Production Models, Individual Speech Peculiarities
Natural Language Texts or Sentences Generator
Phonetics-Acoustic Speech Model Signals Generator
Subject Area 1
Sub-ject Area 2
Subject Area 3
...
...
External World Model
Connected-Speech Signal to be Recognized
Canonical Form
NLTS-Generator
Text or Sentence
PASMS-Generator
Connected Speech Model Signal
Comparator
Natural Language 1 Knowledge (vocabulary, word-building, morphology, phraseology, syntax)
Knowledge about Phonetics, Phonology and Prosody of Natural Language 1, Speech Production Models, Individual Speech Peculiarities

Figure 2. The generalised structure of the Dictation /Translation Machine.

a PASMS-generator. This block takes into account all knowledge as for phonetic and phonological peculiarities of the concrete natural language, all knowledge about a speech production including vocal tract and its excitation source models, all the same about such phenomena as sound coarticulation, its reduction, nonlinear change of rate and intensity of pronouncing, prosodic features and so on. To this block there are introduced the data about individual voice peculiarities in the form of so-called Speech Speaker file (passport) that is the SS-file.

Receiving text or sentence from its NLTS-block the PASMS-generator synthesises all possible connected-speech model signals, which correspond to this text or sentence and simulates the individual speaker voice, which is known for the DTM.

Both generators NLTS and PASMS matched to the language form of the concrete natural language channel that serves all speakers-users of the DTM.

In the frame of the generative model the automatic spoken translation and dictation problems are formulated and solved by the next way.

Simultaneously with the on-line speech signal input all possible continuous-speech model signals are generated by generative model using and compared with the speech signals under analysis. These model signals, under all their varieties, correspond to the appointed texts and sentences, which in their turn, as it follows to the DTM-machine description, express the definite meaning specified by the canonical form. Going through all possible connected-speech model signals and thus all permissible texts and sentences and all possible canonical forms of the meanings to be transmitted let us find the following model signal that has the best (in the definite sense) similarity with the recognisable signal. Then analysing this best connected-speech model signal let us fix to what words sequence that is text or sentence it corresponds and together what meanings or its canonical form is transmitted by this model signal and therefore by the recognisable signal. Since the sought out by this way words sequence is grammatically and semantically correct the latter can be printed if we deal with the dictation machine. Further just as the found words sequence and canonical form express the result of automatic speech recognition and understanding so appealing with this interpretation to the NLTS-generator of the another language it is possible to receive the speech-to-text translation result at its output. Similarly delivering this last speech-to-text translation result to the suitable PASMS-generator input it is generated the connected-speech model signal that is the final result of the full spoken translation from one language to another.

It is appeared some questions as for a realisation of the generative DTM. May be the first of all that pay attention are both sorting out and comparison of connected-speech model signals with recognisable one. It's clear neither today nor in the future it will be discovered such a computer which will be able to realise the full sorting out of various model signals the set of which sufficiently good approximate the real world of speech signals. That is why here are meant only the variants directed sorting out procedures when the current results of model and recognisable signals comparison are used in further speech model signals generative process for the narrowing of the subset of model signals, respectively texts and canonical forms, that pretend to the “recognition response” just so to guarantee not to lose the optimal solution. Under certain conditions this directed sorting out is attained by the mathematical programming methods particularly by dynamic programming (DP). In Figure 2 the variants directed sorting out idea are shown by the feedback arrows that go from the channel *Comparator* (C) to all levels of the DTM hierarchy (H).

External world, NLTS- and PASMS-models designing problems for every language are not less sophisticated. What ever perfect these models are but as for their realisations certainly the two requirements to these models are dominant. These are economical description (specification) of them and the possibility of fast directed sorting out the best variant. These two conflicting requirements are succeeded in satisfying by using of both the stochastic generative grammars in speech model signal generation or composition (C) and the dynamic programming method in the directed sorting out and comparison [3—4].

Don’t going into EW-models designing survey let us describe one variant of the DTM realisation regarding all external world as a simple association of all subject areas.

For each subject area let us define all possible sentences with the aid of a semantic network. All conceivable sentences are divided into categories on the basis of transmitted meaning. The following categories may apply to the information desk of an airport: questions related to flight arrival; the same to flight departure; seat availability; itinerary; location of services, etc. Each category or type of meaning corresponds to its own set of sentences types. The sentence type is the construction that economically specifies a set of sentences that are been obtained from one sentence, by admissible substitutions and inversions of separate words or phrases. A basic element of a sentences type is a subdictionary.

All types of sentences of one meaning are easy to specify using list structure languages, such as LISP. Subdictionaries in sentence types are named and are distinguished as basic. An equivalent means of specifying all possible sentences in the dialogue language is the direct semantic network (DSN). DSN may be constructed by relying upon the types of meanings and the types of sentences. At the same time it is desirable to try to minimize the number of states on the DSN. Any sentence presented may be checked for validity in the DSN, and if valid, can have its meaning expressed in specified canonical form with the aid of the directed semantic network (for example, specified type of presented meaning, type of sentence, names of entities, etc.).

It is possible to propose various approaches to using DSN in the DTM. The first method consists in acoustic elaboration of the DSN. For this, each word in DSN is replaced with the graph (as in HCDP- or HMM-approach), which generates all possible speech signal prototypes. Now the problem of recognising and understanding of connected-speech signal can be formulated as a problem of finding, for the recognisable signal, the most similar reference connected-speech signal from among the set of all prototype signals generated by the connected-speech DSN, and as a problem of analysing (examining) the latter for determination of a series of words and a canonical form of the meaning carried by speech signal [3–4].

4. Multi-level multi-decision model

The second way to create the DTM is multi-level multi-decision model. It is also hierarchically organised. Here significant decisions at all levels of a speech signal processing hierarchy are introduced. For example at the first level the $N1>>1$ best results of speech signal partition into segments corresponding to phonemes are being found. Then at the second level this multi-decision segmentation is being transformed into the $N2>>1$ best phoneme sequence recognition results under free phoneme order. This is a generalised free phoneme recogniser level. At the third level under free word order the $N3>>1$ best word sequences are being found. That is a generalized free word recogniser, and so on. At the highest level there are the $Nh>>1$ best understanding results and the best one is being chosen finally [3-6].

The structure of the multi-level multi-decision dictation/translation machine is shown in Figure 3.

To illustrate how to deal with the multi-level multi-decision model let us consider the two-level multi-decision dictation/translation machine.

At the first level, the generalised problem of connected speech recognition is solved, which consists in that, starting with the assumptions of free word order $N>>1$ different word sentences, arranged in decreasing order of similarity, are found that most resemble the signal under processing. Then, at the second level, these N sequences are analysed in order until a sequence of words is found that coincides with one synthesised by the NLTS-generator or belongs to one of the types of sentences and correspondingly to one of the types of meaning. This sequence of words is declared to be the result of recognition, and the canonical form of the transmitted meaning is composed with the aid of the DSN or with the aid of the types of meaning and the types of sentences [3-6].

Now let us consider the HCDP phoneme-by-phoneme recognition problem for continuous speech, its generalisation and applications in ASR [3, 8–9]. Simultaneously INTAS cooperation proposal will be discussed.

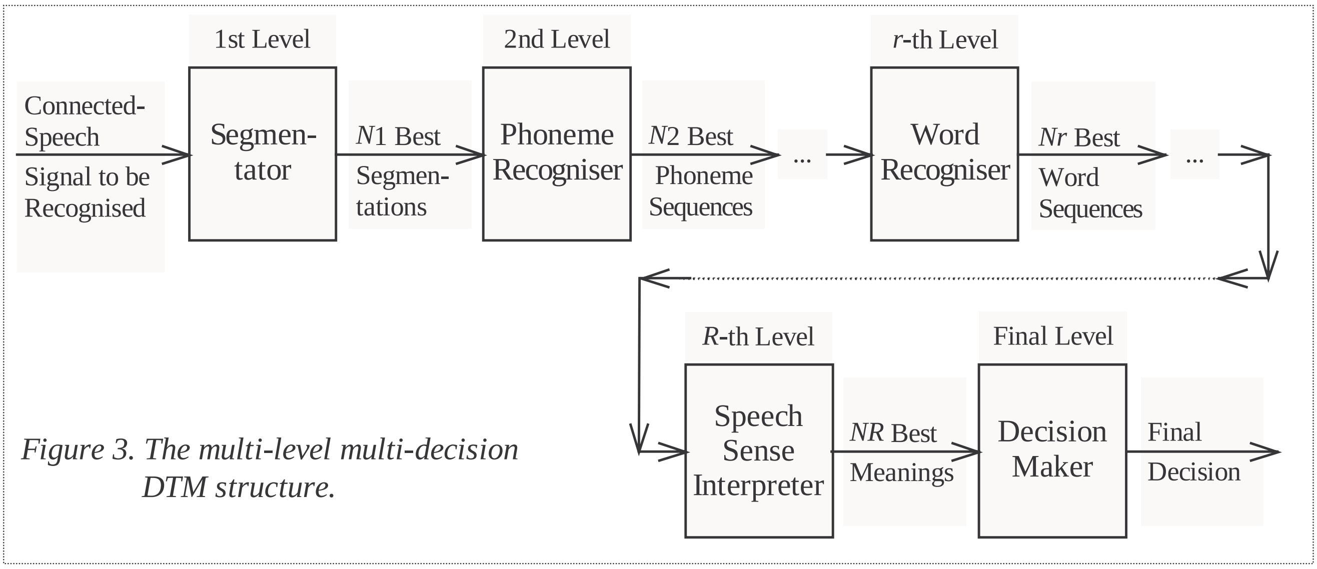

Figure 3. The multi-level multi-decision DTM structure.

5. Phoneme-by-phoneme ASR problem

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.

Although 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 as well as between research groups, especially from INTAS cooperation.

To improve this approach it was proposed to introduce significant decisions in phoneme recognition procedures [3, 8–9]. The next step consists in making improvements to used generative automata grammars, for example instead of phoneme-diphones speech model [8] to put into operation a phoneme-threephones one [9].

Here 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 Figure 4. 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.

6. Phoneme-by-Phoneme Recognition in Continuous Speech

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.

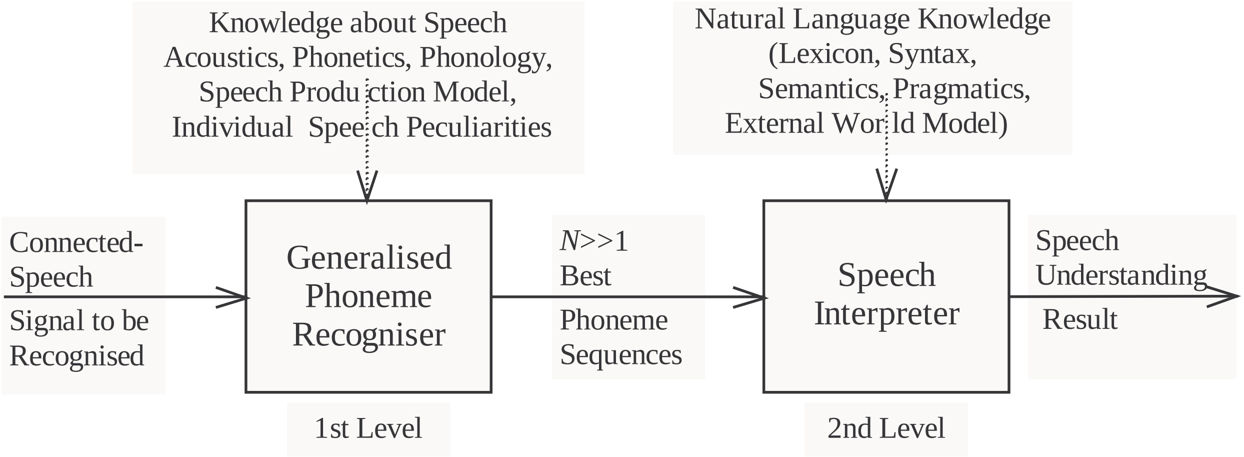
Knowledge about Speech Acoustics, Phonetics, Phonology, Speech Production Model, Individual Speech Peculiarities
Natural Language Knowledge (Lexicon, Syntax, Semantics, Pragmatics, External World Model)
Connected-Speech
Signal to be Recognised
Generalised Phoneme Recogniser
N>>1 Best
Phoneme Sequences
Speech Interpreter
Speech Understanding
Result
1st Level
2nd Level

Figure 4. Two-Level Speech Understanding System Structure.

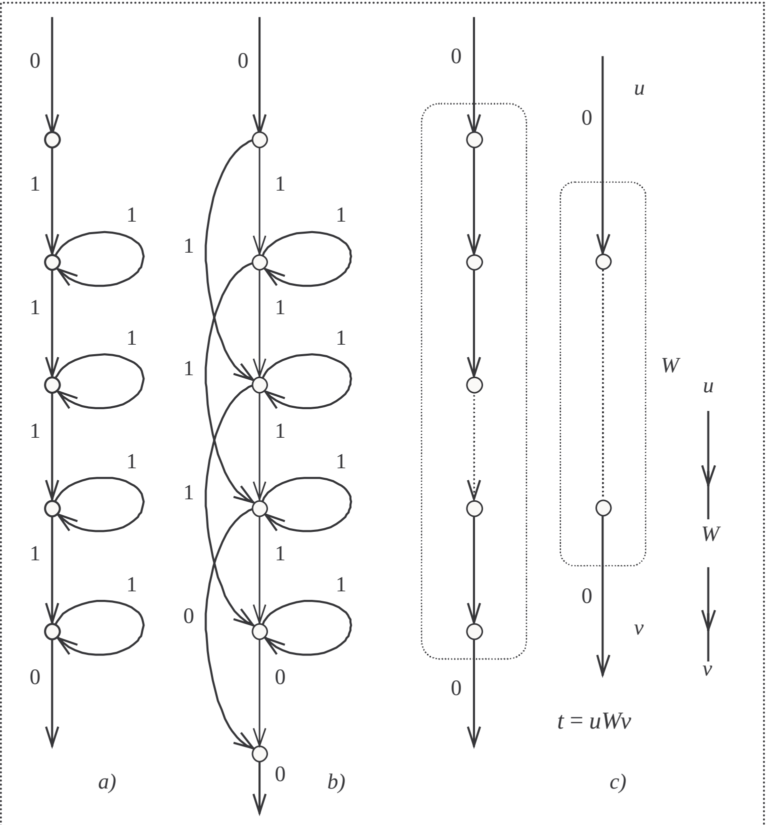

Figure 5. Generative grammars (graphs) for the phoneme-threephone: a) no microelement omission; b) no two microelements running omission; 3) schematic notes of the PT graph $t = uWv$.

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 [1].

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.

This mentioned generative grammar for free phoneme sequences will be given under phoneme-threephones (PT) interpretation.

Let be given the finite set K of the phonemes $k\in K$. The phoneme alphabet includes the phoneme-pause #. For example, in the Ukrainian phoneme alphabet 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 softness], H, H', ZH, ZH', Z, Z', J, L, L', M, M', N, N', R, R', F, F', KH, KH', SH, SH', #} $\equiv K^{\mathrm{st}}\subset K$, which change their duration, and transitive phonemes $k\in$ {B, B', G, G', D, D', K, K', P, P', T, T'} $\equiv K^{\mathrm{tr}}\subset K$.

Generally speaking, natural language allows totally $|K|^3$ PT but hereafter there are considered only about 2,000–3,000 basic PTs $t\in T$, which approximate all possible PTs. We remind that 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 or non-terminal symbols for PT t, respectively. So, the PT t=uWv is the phoneme W that is considered under influence of neighbouring phonemes u and v. They are the first u that precedes W and the second v that follows W.

Obviously, only PTs t_1=uWv and t_2=wVz are alowable for connection via Wv and wV.

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 basic PT $t\in T$. The PT t is specified by its acoustical transcription in the alphabet J: $t = (j_{t1},\ j_{t2},..., j_{ts},..., j_{tq(t)})$, where s indicates the ordinal place in the transcription 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 [3, 8–9]. For each speaker they form the basis of the so-called Speaker Voice File (Passport) [7].

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

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

where $r_0 = m$, $r_{s\text{-}1}< r_s$, $r_{q(t)}= n$. 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 Figure 5a. 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 Figure 5b forbids removing of more than two microphonemes running. Schematic notes for PT graph t=uWv, $u,W,v\in K$ are given in Figure

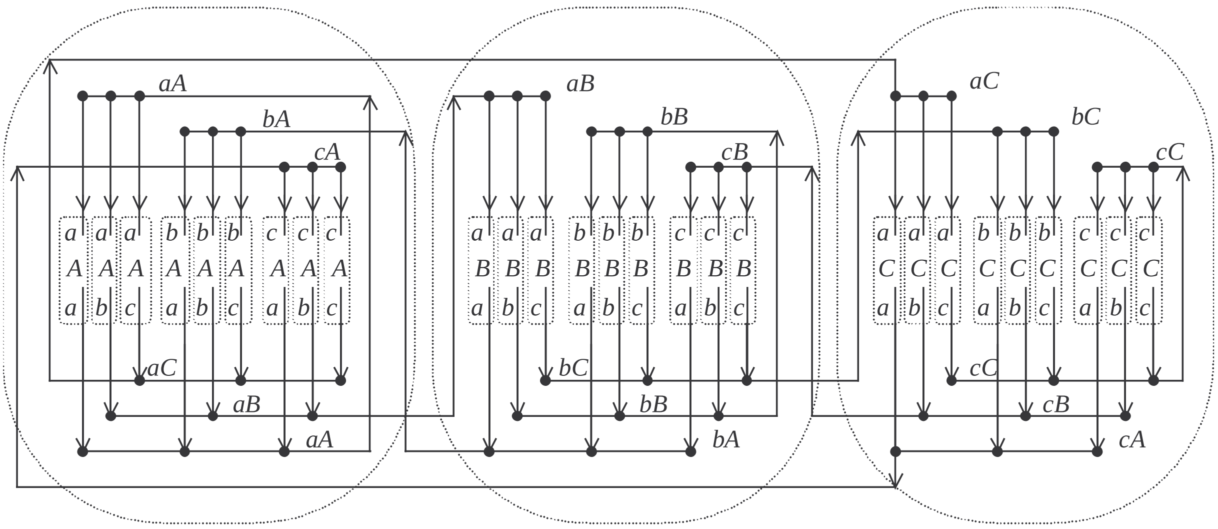

Figure 6. Common phoneme graph for phoneme-by-phoneme continuous speech recognition.

5c, 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 allowable to connect PTs $t_1=uWv$ and $t_2=wVz$ into phoneme sequence so that the output pair name Wv of preceding PT t_1 coincides with the input phoneme pair name wV of the following t_2.

Going such a way it will be received a common phoneme graph (CPG) for continuous speech signal generation. The full CPG for three phoneme alphabet $K=\{A, B, C\}$ is shown in Figure 6.

For each phoneme $W\in K$ it is corresponded a block of all $|K|^2$ PTs uWv, where $u,v\in K$. In Figure 6 the blocks are denoted by dotted line. Each block of phoneme W has $|K|$ input buses uW, $u\in K$ and $|K|$ output buses wZ, $Z\in K$. Output buses wZ are connected with input buses under the same name.

Additionally, it is distinguished the input state $s=u$, the output state $s=v$ and internal states s for each PT $t=uWv$, $u,W,v\in K$. One of phonemes is associated with the phoneme-pause #. It means that the block of phoneme # has an input bus ## being started at $i=0$. Let us also introduce the overall enumeration of states within each PT on 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):

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

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

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 input bus s=## to any 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 speech 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 by DP recurrent formulae [9].

For the phoneme sequence recognition response forming it is sufficient to remain in the memory the 3-le array $F_i(\alpha)$, $n_i(\alpha)$, $t_i(\alpha)$, $\alpha=wV$, $w,V\in K$, $i=1:l$ and to avail oneself of the extracting algorithm [9].

When the generalised phoneme-by-phoneme recognition problem it is necessary to locate in memory the N-le not of 3-le but 4-le ($F_i^r(\alpha)$, $n_i^r(\alpha)$, $\beta_i^r(\alpha)$, $z_i^r(\alpha)$), $i=1:l$, $r=1:N$ for all $\alpha=wV$, $w,V\in K$, where $z_i^r(\alpha)$ indicates the r-th place in the previous N-le, and then to use a little complicated extracting algorithm [9].

7. The INTAS proposal summary

The INTAS project titled “Language independent model for ASR advanced research” is proposed. It is based on investigation of generalised phoneme-by-phoneme recognition models for automatic transformation of speech signal into universal phonetic text, acceptable for processing, interpretation and undestanding in any language.

Movements for a speech production model and their clusterisation are studied, and universal language independent phonetic alphabet (ULIPA) for marking and classification of these movements is established.

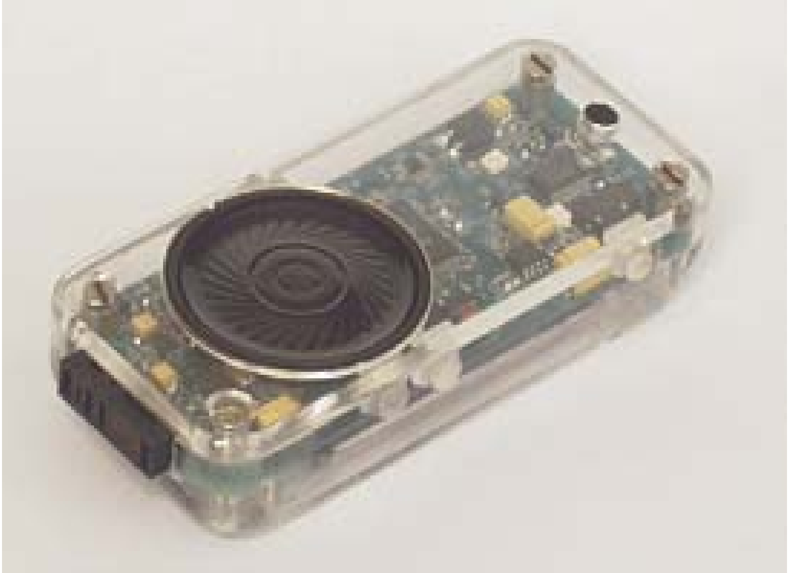

Figure 7. Digital recorder totally controlled by voice

Different speech signal feature descriptions and speech signal similarity measures are considered. Advances to their generalisation and unification as well as respective associations with articulation movements (gestures) and their comparison are performed.

The language independent ASR problem is considered as a discovering problem for hidden sequence of typical articulators movements, respectively for hidden sequence of universal phonetic ULIPA symbols. For this discovering mentioned researches are being carried out, and generalised robust hidden models for speech signal presentation by one or N>1 best sequences of universal phonetic ULIPA symbols are researched.

Existing speech signal databases are adapted for Automatic Language Independent Speech Recognition (ALISR) training. Supervised and unsupervised procedures for automatic speech signal segmentation, comparison, clustering, labelling in ULIPA and other alphabets as well as all kinds of ALISR training procedures are studied.

Generalised robust hidden models for automatic transformation of speech signal into ULIPA or like another language independent phonetic text are researched. Connections between automatic unsupervised speech labelled texts, ULIPA phonetic texts and different natural language phonetic and orthographic texts are established.

Methodology and technology for designing of language independent speech recognition and understanding systems are created. Our knowledge as concerns speaker and natural language speech signal variability is being extended. Some regularities in speech signals for different languages are being found out. Software ALISR prototypes, demonstrations, papers, presentations are expected. East and West scientific collaboration is advanced.

8. References

[1] Final Report on the UNESCO contract SC/RP 261060.8 “Development of Multilingual (including English and Russian Languages) Speech Dialogue System for a Microcomputer”. – Kiev–Paris, 1986, 97 p.

[2] INTAS project 93–2119 “Extension of ELSNET to the NIS”.

[3] T.K. Vintsiuk. Analysis, Recognition and Understanding of Speech Signals. – Kiev: Naukova Dumka, 1987, 264 p, in Russian.

[4] Taras K. Vintsiuk. *Two Approaches to Create a Dictation/Translation Machine.* – Proc. of the Workshop SPECOM’97, Cluj-Napoca, 1997, pp 1-6.

[5] Taras K. Vintsiuk. *Generalized Problem for Automatic Phoneme Recognition.* – Proc. of the Workshop SPECOM’97, Cluj-Napoca, 1997, pp 115-118.

[6] Taras K. Vintsiuk. *Significant Speech Undestanding Algorithm for a Dictation/Translation Machine.* – Proc. of the Workshop SPECOM’98, St.-Petersburg, 1998, pp 197-202.

[7] Taras K. Vintsiuk, Mykola M.Sazhok. *Speaker Voice Passport for a Spoken Dialogue System.* – Proc. of the Workshop SPECOM’98, St.-Petersburg, 1998, pp 275-278.

[8] Vintsiuk T.K. Avtomatika Vol. 6, pp. 40–49 (1972), Vol. 1, pp. 63–72 (1973).

[9] T.K. Vintsiuk *Generative Phoneme-Threephone Model for ASR.* – Proceedings of the International Conference “Text, Speech and Dialogue”, 2001, LNAI 2166 , pp 201-207.

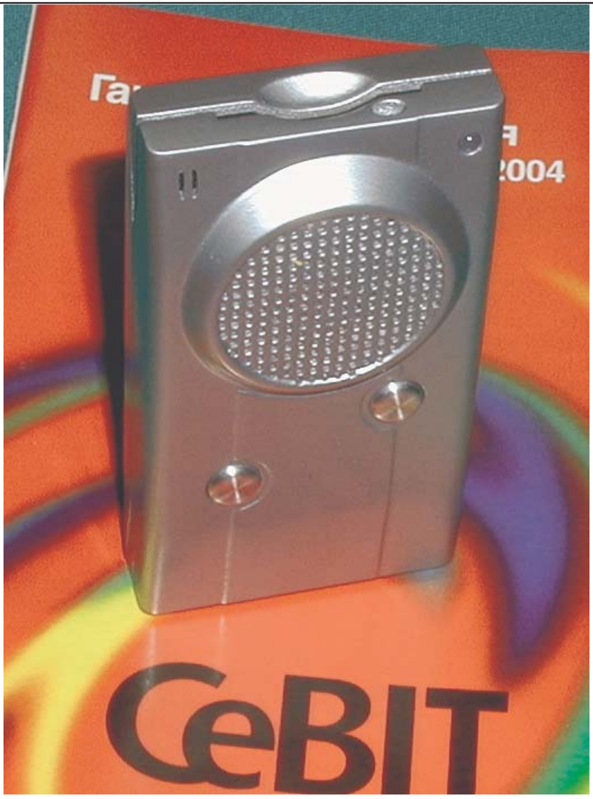

Figure 8. Spoken interpreter containing about 300 words and phrases