Multi-Level Multi-Decision Models in ASR

Taras Vintsiuk, Mykola Sazhok

Speech Science and Technology Department
Int. Research/Training Center for IT&S – IRTC, Kyiv, Ukraine
{vintsiuk, mykola}@uasoiro.org.ua

Abstract

Multi-Level Multi-Decision Models for Automatic Speech Recognition is discussed. It is hierarchically organized. Here there are not used the generative grammars for model speech signal synthesis as a feedback in speech recognition process. Instead of the latter significant decisions, but under simplified conditions, at all levels of a speech signal processing hierarchy are introduced. The 3-level model with phoneme recognizer, word recognizer and continuous speech interpreter is proposed. Experimental results for the 3-level model are given and problems to be solved are discussed.

1. Introduction

At present the investigators who acknowledge the possibility of phoneme speech understanding have two different approaches to the problem [1],[2]. The followers of the first approach assume that continuous speech must firstly be recognized as a phoneme/syllable sequence, and then this phoneme sequence must be recognized and understood as word sequence and meaning to be transmitted by a speech signal, respectively. In contrast, the followers of the second approach assume that understanding needs not precede phoneme nor word recognition, and if phoneme recognition is nevertheless carried out, then it must be simultaneous with speech understanding. Moreover, the phoneme recognition must not be rigid but controlled in such a way to yield the best result of understanding.

It is easy to see that the first approach is erroneous, since the best method of finding of phonemes to be transmitted is both to recognize and to understand a speech signal. Only after that it will be possible to determine rigorously the phoneme and word sequence corresponding to the speech signal, i.e., phoneme recognition and speech understanding must be interrelated. Therefore the only acceptable approach is the second.

But this second approach is very complicated because it makes to operate simultaneously with all the knowledge about human being—natural language—speech phenomena. Moreover it complicates the job distribution between specialists in acoustics, phonetics, linguistics and informatics.

These lacks are shown feeble in the first approach. That is why to improve the latter it is proposed to introduce significant decisions in phoneme recognition procedures.

In this paper we propose a so-called generalized phoneme recognition problem for the three-level speech recognition system. The structure of this system is shown in Figure 1. It is consists of three parts. These are Generalized Phoneme Recognizer, Generalized Word Recognizer and Continuous Speech Interpreter.

A generalized phoneme recognition problem means that under free phoneme order it is being found the *N*>>1 best

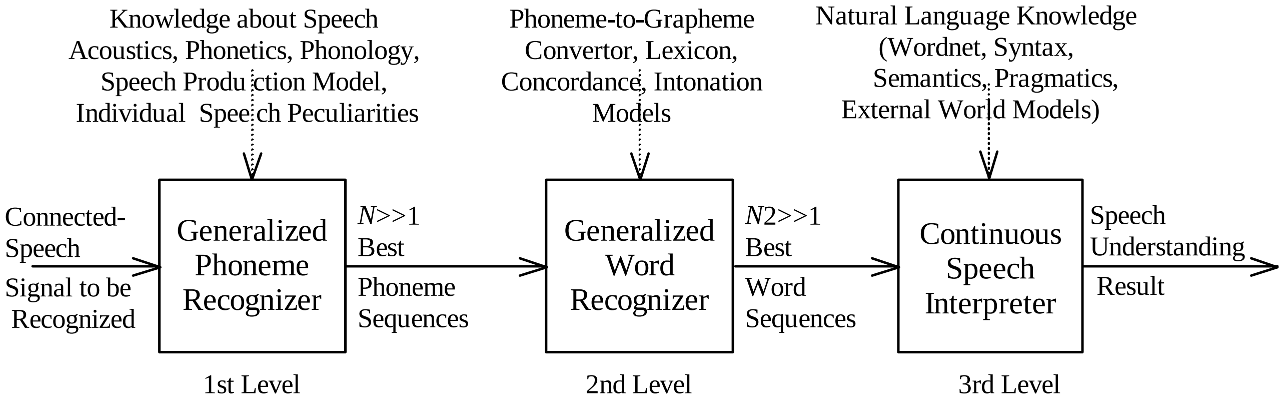

Knowledge about Speech Acoustics, Phonetics, Phonology, Speech Production Model, Individual Speech Peculiarities
Phoneme-to-Grapheme Convertor, Lexicon, Concordance, Intonation Models
Natural Language Knowledge (Wordnet, Syntax, Semantics, Pragmatics, External World Models)
Connected-Speech Signal to be Recognized
Generalized Phoneme Recognizer
N>>1 Best Phoneme Sequences
Generalized Word Recognizer
N2>>1 Best Word Sequences
Continuous Speech Interpreter
Speech Understanding Result
1st Level
2nd Level
3rd Level

Figure 1. Three-Level Speech Recognition System Structure.

phoneme sequence recognition responses. Then a Generalized Word Recognizer analyses these phoneme sequences in order to generate $N2$>>1 possible word sequences. By these word sequences a Speech Interpreter makes a decision about the speech understanding response via Natural Language Knowledge.

2. Generalized Phoneme Recognizer

2.1. General idea

The general idea is, taking into account inertial properties of articulation apparatus and language phonetics only, to construct some phoneme generative automata grammar which can synthesize 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.

To take into account the fact of co-articulation in [3] we considered a phoneme-threephone model. But here we deem a monophone model believing that multi-decision and multi-level factors as well as GMM will compensate this simplification.

The problem of directed synthesis, sorting out and formation of a phoneme sequence recognition response is solved by using the special computational scheme of dynamic programming, in which (for a substantial reduction in memory and calculation requirements) the concepts of potentially optimal both index and phoneme are used [1],[3].

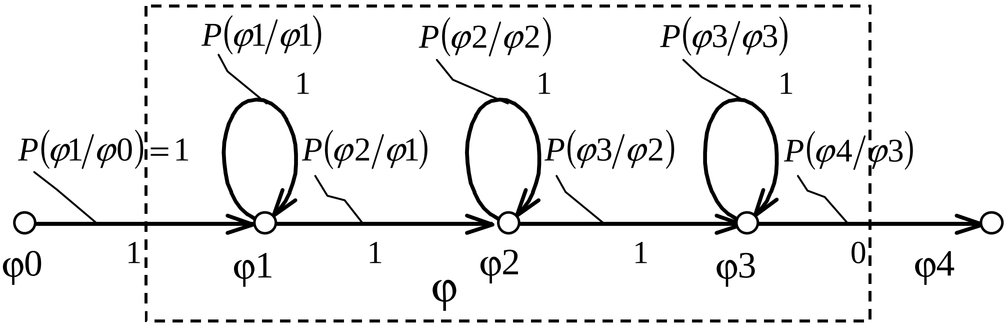
At first, the phoneme-by-phoneme continuous speech recognition problem will be analyzed. Then this research will be generalized for N>>1 best phoneme sequences.

2.2. General free phoneme sequence generative grammar

This mentioned generative grammar for free phoneme sequences will be given under the monophone interpretation unlike the diphone/threephone one in [1, 3].

From now on we assume that besides phoneme alphabet we have such knowledge:

Each phoneme φ from the alphabet Φ of basic phonemes (for Ukrainian, |Φ| = 55) is modeled with a stochastic generative grammar like in Fig. 2 consisting of 5 states: φ0 and φ4 are start and end states respectively; φ1, φ2, φ3 are the 3 states simulating 3 hypothetical phase of the phoneme φ dynamics. The parameters of the phoneme generative grammar are $P(\varphi 1/\varphi 1)$, $P(\varphi 2/\varphi 2)$, $P(\varphi 3/\varphi 3)$. Also we suggest known distributions $P(x/\varphi t)$, $x \in \Xi$, $t = 1,3$, where Ξ is a space of observed elements-vectors. These parameters for all phonemes make a so-called Speaker Voice Passport for a person or a cooperative of persons and are estimated during the training or self-training procedure [4].



$$P(\varphi 1/\varphi 1)+P(\varphi 2/\varphi 1)=1 \qquad P(\varphi 2/\varphi 2)+P(\varphi 3/\varphi 2)=1 \qquad P(\varphi 3/\varphi 3)+P(\varphi 4/\varphi 3)=1$$

$$P(x/\varphi t),\, x\in \Xi,\, t=1:3$$

Figure 2. Generative grammar for the phoneme φ composed of three phases φt, t = 1,2,3.

The probability that a segment $X_{\mu\nu} = (x_{\mu+1}, x_{\mu+2}, \ldots, x_i, \ldots, x_\nu)$, $0 \le \mu < \nu \le l$ of the observed signal $X_{0l} = (x_1, x_2, \ldots, x_i, \ldots, x_l)$ with length l belongs to the phoneme φ might be written as:

$$P(X_{\mu\nu}/\varphi) = \max_{(w_1,w_2)} \left\{\left[(P(\varphi 1/\varphi 1))^{w_1-\mu-1}(1-P(\varphi 1/\varphi 1))\prod_{i=\mu+1}^{w_1} P(x_i/\varphi 1)\right]\times \right.$$
$$\times\left[(P(\varphi 2/\varphi 2))^{w_2-w_1-1}(1-P(\varphi 2/\varphi 2))\prod_{i=w_1+1}^{w_1} P(x_i/\varphi 2)\right]\times \qquad (1)$$
$$\left.\times\left[(P(\varphi 3/\varphi 3))^{\nu-w_2-1}(1-P(\varphi 3/\varphi 3))\prod_{i=w_2+1}^{\nu} P(x_i/\varphi 2)\right]\right\}$$

where (w_1, w_2): $\mu < w_1 < w_2 < \nu$ are the bounds of phoneme phases.

Uniting graphs of phoneme generative grammars under the free-phoneme order condition we receive a common phoneme graph (CPG). The full CPG for the 6 phoneme alphabet $\Phi = \{\varphi: \varphi=1,2,3,4,5,6\}$ is shown in Figure 3. The transitions between states are doing in accordance to arrows and during 0 or 1 discrete time steps. Each discrete step i is associated with observation of x_i.

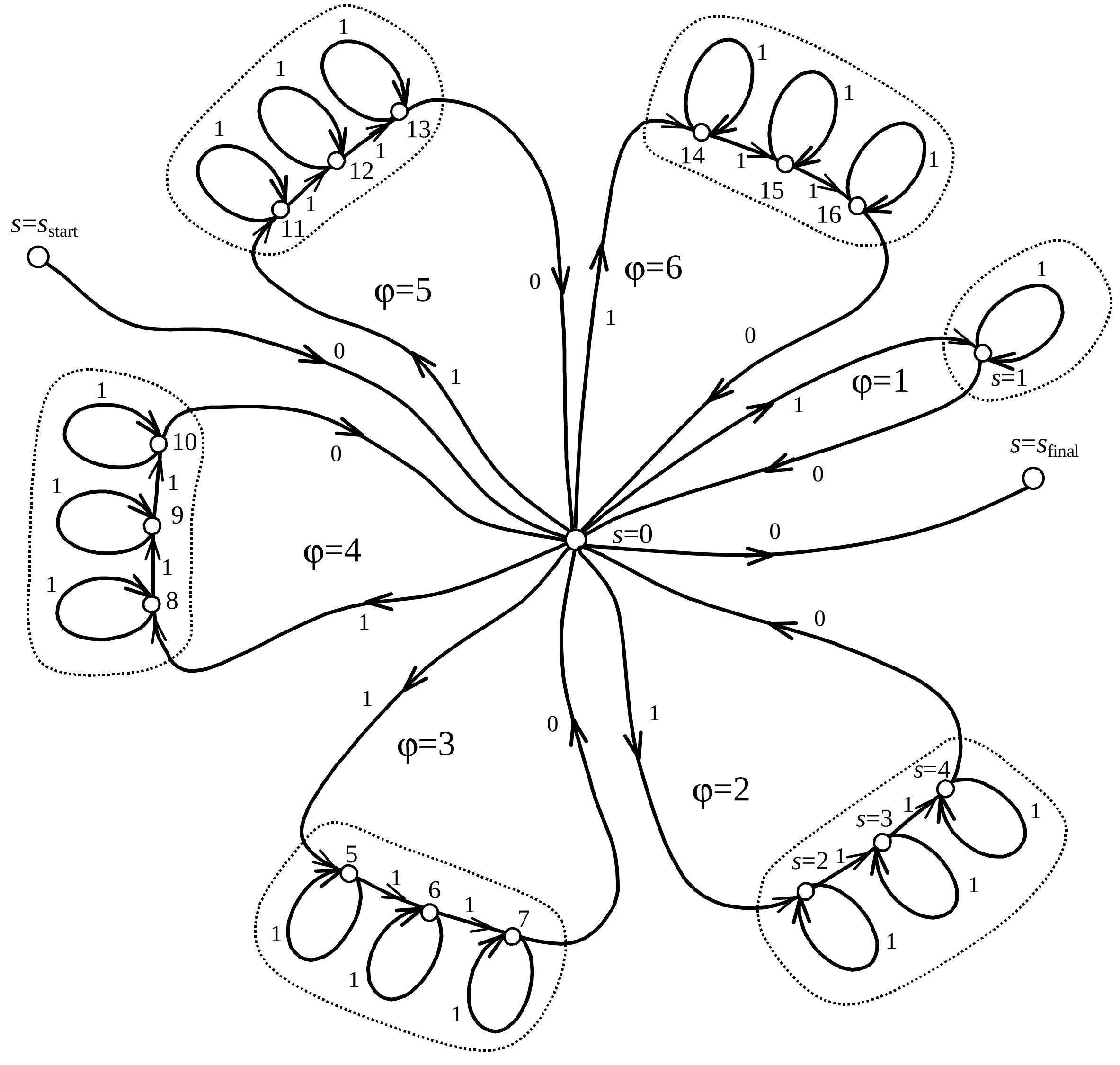


Figure 3. Free phoneme sequences generative grammar for continuous speech.

Thus, accordingly to CPG the probability of the observed speech signal $X_{0l} = (x_1, x_2, \dots, x_i, \dots, x_l)$ where l is length of the observed signal under condition of a hidden phoneme sequence $\Phi_{0Q*} = (\varphi_1, \varphi_2, \dots, \varphi_u, \dots, \varphi_{Q*})$ might be computed by the formula:

$$P(X_{0l} / \varphi_1, \varphi_2, \dots, \varphi_u, \dots, \varphi_{Q*}) = \max_{\{\mu_u, Q\}} \prod_{u=1}^{Q} P(X_{\mu_u \mu_{u+1}} / \varphi_u) \quad (2)$$

where probabilities $P(X_{\mu_u \mu_{u+1}} / \varphi_u)$ are calculated by (1) and $0 = \mu_0 < \mu_1 < \dots \ \mu_u < \ \dots < \mu_Q = l$ are the phoneme bounds and $Q*$ is a quantity phoneme samples in the hidden sequence.

2.3. Phoneme sequence recognition algorithm

As far our task satisfies the Bellman's Optimality Principle, we apply it in the algorithm formulation. Figure 4 explains this on example of 4 phonemes and a phoneme-pause φ=1. The graph in Figure 4 allows for computing the probability (2). States s are enumerated in order of their passing. The basic state s=0 and one state for a pause s=1 are introduced. During transition to a state s in time moment i an observed element x_i is associated with the 1st, 2nd or 3rd state s of the respective phoneme-phase φt(s).

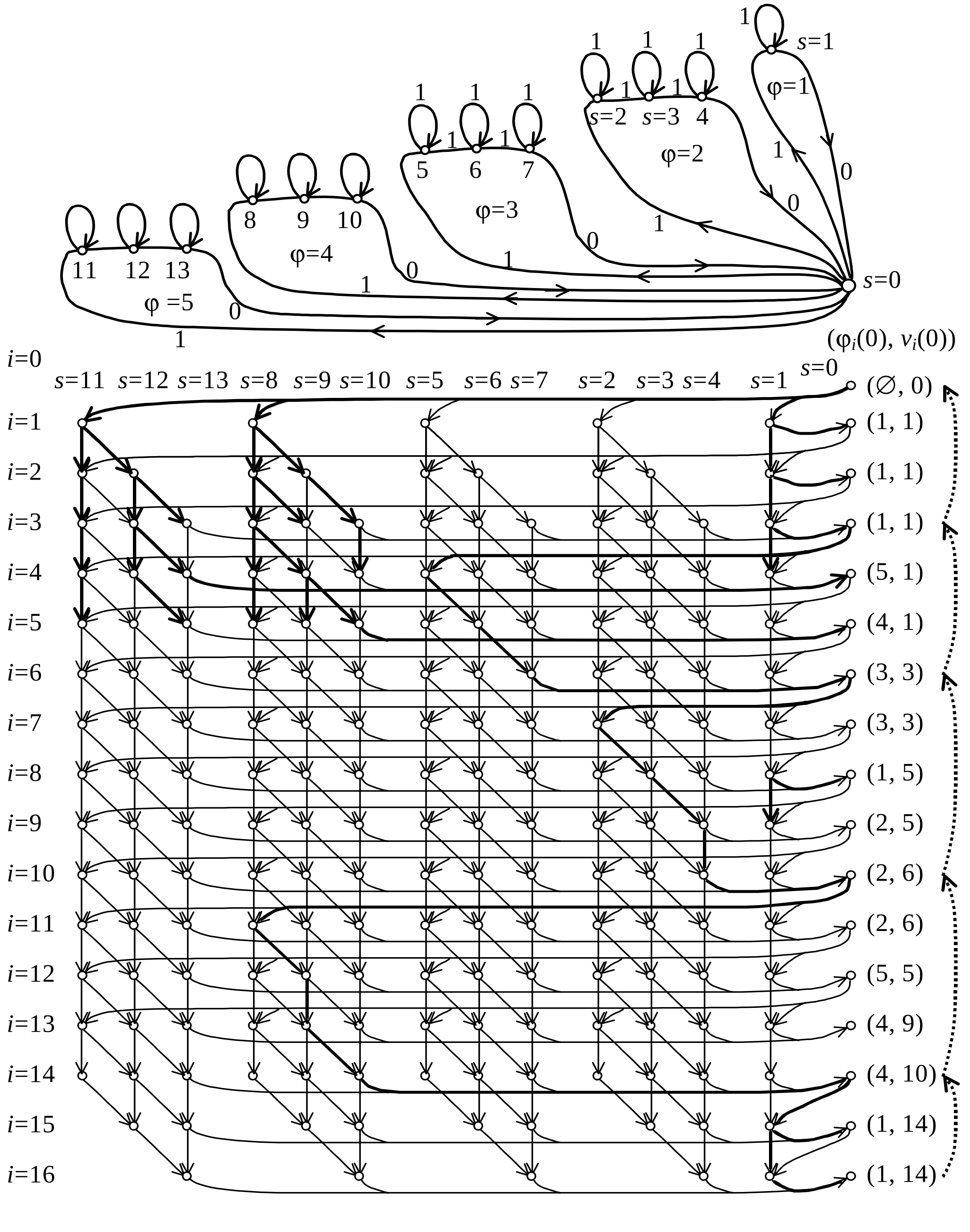


Figure 4. Free phoneme sequences generative grammar for continuous speech composed of phonemes from the alphabet Φ, |Φ|=5 including a phoneme-pause, shown on the hypothetical speech control sample consisting of 16 elements. Bold lines (both regular and dashed) show the optimal trajectory and recognition response forming by the array ($\varphi_i(0)$, $v_i(0)$). The potentially optimal trajectories that do not reach the final state are partially shown with bold lines.

Let be designated by $\Omega_i(s)$ a set of continuous speech trajectories of duration i which are permissible in the CPG and which are results of movements from state s=0 to state s in 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 observed segment $X_{0i} = (x_1, x_2,..., x_i)$, and by $v_i(s)$ the potentially optimal beginning of the last phoneme $\varphi_i(s)$ in the best phoneme sequence for $\Omega_i(s)$. Then $\varphi_i(0)$ will be the name of the phoneme which potentially finished in the state s=0 at the moment i and $v_i(0)$ will be its potentially time beginning in the state s=0 [1, 3].

We will distinguish: the state s=0 is a main state, s=1 is a phoneme-pause state; s=2,3,4, s=5,6,7, s=8,9,10 and s=11,12,13 represent 1st, 2nd and 3rd phase of 1st, 2nd, 3rd and 4th phoneme, respectively. Let be denoted by IN a set of all CPG internal states that is $IN = \{s:\ s = 3,6,9,12\}$, $INP = \{s:\ s = 1,2,5,8,11\}$ is a set of all CPG input states and $OUT = \{s:\ s = 1,4,7,10,13\}$ is a set of output states. Therefore, a phoneme-pause state s=1 is associated with both INP and OUT sets. Union of all internal and output states makes an OWN set.

Hereafter, a notation $\varphi t(s)$ means a phase t of phoneme φ considered under state s, e.g. $\varphi 2(3)$ means a phase 2 for φ=2.

Let $F_u(s)$, $\varphi_u(s)$, $v_u(s)$ have been calculated for all states s and for all time steps $u<i$ which precede i. Then after the next observed element x_i appearance simultaneously for all states s are calculated new values $F_i(s)$, $\varphi_i(s)$, $v_i(s)$ in the following order a) – c):

a) for all states $s\in OWN$ and for all φ:

$$\begin{aligned} F_i(s) = [\max\{&F_{i-1}(s-1)P(\varphi t(s)/\varphi t(s-1)),\\ &F_{i-1}(s)P(\varphi t(s)/\varphi t(s))\}]P(x_i/\varphi t(s)), \end{aligned} \tag{3}$$

$$v_i(s) = v_{i-1}\left(s - I\begin{pmatrix} F_{i-1}(s-1)P(\varphi t(s)/\varphi t(s-1)), \\ F_{i-1}(s)P(\varphi t(s)/\varphi t(s)) \end{pmatrix}\right) \tag{4}$$

where $I\begin{pmatrix} a, \\ b \end{pmatrix} = \begin{cases} 1, & if\ a > b; \\ 0, & if\ a \le b. \end{cases}$

b) for all input states $s\in INP$ (they are s=1,2,5,8,11 in Fig. 4) of all phonemes φ:

$$F_i(s) = [\max\{F_{i-1}(0), F_{i-1}(s)P(\varphi 1(s)/\varphi 1(s))\}]P(x_i/\varphi 1(s)), \tag{5}$$

$$v_i(s) = \begin{cases} i-1, & if\ F_{i-1}(0) > F_{i-1}(s)P(\varphi 1(s)/\varphi 1(s)); \\ v_{i-1}(s), & if\ F_{i-1}(s)P(\varphi 1(s)/\varphi 1(s)) \le F_{i-1}(0). \end{cases} \tag{6}$$

If in (6) v_i(s) is assigned to i–1 then at the time moment i a new phoneme $\varphi(s)$ has potentially started;

c) for the main state s=0 where phoneme potentially starts and potentially ends:

$$F_i(0) = \max_{s\in OUT}[F_i(s)(1 - P(\varphi t(s)/\varphi t(s)))]; \tag{7}$$

$$s_i(0) = \operatorname*{argmax}_{s\in OUT}[F_i(s)(1 - P(\varphi t(s)/\varphi t(s)))]; \tag{8}$$

$$\varphi_i(0) = \varphi(s_i(0)); \tag{9}$$

$$v_i(0) = v_i(s_i(0)). \tag{10}$$

In (7)–(10) $\varphi_i(0)$ is phoneme that potentially ended in time moment i and $v_i(0)$ is a potentially optimal moment of its start.

For the phoneme sequence recognition response forming it is sufficient to retain in memory the 3-le array of ($F_i(0)$, $\varphi_i(0)$, $v_i(0)$) for i=0:l.

Since we imply a continuous speech signal X_{0l} begins and ends with a phoneme-pause the phoneme sequence recognition response is formed by the following extracting algorithm:

Let be $v_1^* = l$ and φ_1^* is a phoneme-pause. Then for ρ=1,2,3,... there will be extracted $v_{\rho+1}^* = v_{v_\rho^*}(0)$ and $\varphi_{\rho+1}^* = \varphi_{v_\rho^*}(0)$ until $v_{\rho+1}^* = 0$ will be reached.

Hence, the phoneme sequence φ_ρ^*, ρ=1,2,3,... will be the phoneme sequence $\varphi_{\rho+1}^* = \varphi_{v_\rho^*}(0)$ recognition response in the opposite direction and v_ρ^*, ρ=1,2,3,... will be the respective phoneme bounds in the signal X_{0l}.

To begin the recognition process (3)—(10) it is assumed that $F_0(0)$=1, $v_0(0)$=0 and $F_0(s)$=0 for all s>0.

2.4. The generalized algorithm

To find N>>1 best phoneme sequences in the signal X_{0l} let us modify the basic algorithm 2.3.

Now for all states s and for any time step i there will be calculated N-let of not a triple but a quadruple

$$\left(F_i^r(s), \varphi_i^r(s), v_i^r(s), k_i^r(s)\right),\ r=1:N \tag{11}$$

which is composed of N best probabilities $F_i^r(s)$ that correspond to N best but different phoneme sequence recognition potential responses for X_{0i}. In (11) the $k_i^r(s)$ is the ordinal number of the quadruple preceding a considered quadruple: $\left(F_i^r(s), \varphi_i^r(s), v_i^r(s), k_i^r(s)\right)$. The quadruples are ordered in a way that $F_i^1(s) \ge F_i^2(s) \ge \ldots \ge F_i^r(s) \ge \ldots$ $\ldots \ge F_i^N(s)$ and all respective $\varphi_i^r(s), r = 1:N$ corresponds to different potential phoneme sequence responses.

Computation is started with the following assumptions: $F_0^1(0) = 1$ and $F_0^r(s) = 0$ for all s>0 and $r\in$[2, N].

Let the N-let (11) has been calculated for all states s and for all time steps $u<i$ which precede i. Then after the next observed element x_i appearance simultaneously for all states s a new N-let (11) is calculated in a way similar to a) – c) in 2.3 with ranging N best decisions by their probability decrease:

a) for all states $s\in$OWN and for all φ the 2N products $F_{i-1}^r(s-1)P(\varphi t(s)/\varphi t(s-1))$, $F_{i-1}^q(s)P(\varphi t(s)/\varphi t(s))$, r,q =1:N are composed. Among these products N best ones multiplied by $P(x_i/\varphi t(s))$ are ranged and stored to $F_i^r(s)$, r =1:N with simultaneous filling the other components $v_i^r(s)$ and $k_i^r(s)$;

b) for all input states $s\in INP$ of all phonemes φ the 2N values $F_{i-1}^r(0)$, $F_{i-1}^q(s)P(\varphi 1(s)/\varphi 1(s))$, r,q =1:N are composed. Among these products N best ones multiplied by $P(x_i/\varphi 1(s))$ are ranged and stored to $F_i^r(s)$, r =1:N with simultaneous filling the other components $v_i^r(s)$ and $k_i^r(s)$.

c) for the main state s=0 where phoneme potentially starts and potentially ends the $|\Phi|N$ products

$F_{i-1}^{r}(s)(1-P(\varphi t(s)/\varphi t(s)))$, r =1:N , $s\in OUT$ are evaluated, ranged and N best of them are stored to $F_i^{r}(0)$ among with the other components $\varphi_i^r(0)$, $v_i^r(0)$ and $k_i^r(0)$.

Let us emphasize that in each N-let at least one of components $\varphi_i^r(0)$, $v_i^r(0)$ and $k_i^r(0)$ should be different.

Now for the generalized phoneme sequence recognition response forming it is necessary to locate in memory the N-let of values (11) for s=0, i=1:l and then to use the a little complicated extraction algorithm.

Actually, the array of $\left(F_i^{r}(0),\varphi_i^r(0),v_i^r(0),k_i^r(0)\right)$ for all i=1:l containing 4Nl values is the 1st level output.

However, the generalized recognition response might be formed. For that it should be used N times an extracting algorithm in 2.3 changing the r from 1 to N and proceeding from that that in the finale N-let the last phoneme is a pause.

Let be $v_1^{*r}=l$ and φ_1^{*r} is a phoneme-pause. Then for ρ=1,2,3,... there will be extracted $v_{\rho+1}^{*r}=v_{v_\rho^{*r}}(0)$ and $\varphi_{\rho+1}^{*r}=\varphi_{v_\rho^{*r}}(0)$ until $v_{\rho+1}^{*r}=0$ is reached.

Hence, the phoneme sequence φ_ρ^{*r}, ρ=1,2,3,... will be the r-th phoneme sequence $\varphi_{\rho+1}^{*r}=\varphi_{v_\rho^{*r}}(0)$ recognition response in the opposite direction and v_ρ^{*r}, ρ=1,2,3,... will be the respective phoneme bounds in the signal X_{0l}.

3. Generalized Word Recognizer

One of the Phoneme Recognizer level result is N>>1 best observed phoneme sequences $\Phi_{0Q^r}^{r}=\left(\varphi_1^r,\varphi_2^r,\ldots,\varphi_u^r,\ldots,\varphi_{Q^r}^r\right)$, r=1:N where Q^r is a length of the r-th observed sequence. Moreover, as the result of the first level, each phoneme observation φ_u^r might be accomplished with information about its duration d_u^r, probability ΔF_u^r and may be other parameters like energy, pitch movement etc.

At the second level Word Recognizer must extract for all $\Phi_{0Q^r}^{r}$, r=1:N total $N$1>>1 hidden phoneme sequences $\Psi_{0Q^{r1}}^{r1}=\left(\psi_1^{r1},\psi_2^{r1},\ldots,\psi_s^{r1},\ldots,\psi_{Q^{r1}}^{r1}\right)$, r1=1:$N$1, $\psi\in\Psi\equiv\Phi$ and associate them with word sequences $J_{0Q^{r2}}^{r2}=\left(j_1^{r2},j_2^{r2},\ldots,j_k^{r2},\ldots,j_{Q^{r2}}^{r2}\right)$, r2=1:$N$2, $N$2>>1 and $j_k^{r2}\in J$ where J is a word dictionary. To avoid loosing the actual word sequence $N$2>>1 recognition responses are taken.

Thus, we interpret observed phoneme subsequences $\Phi_{u_{s-1}u_s}^{r}=\left(\varphi_{u_{s-1}+1}^r,\varphi_{u_{s-1}+2}^r,\ldots,\varphi_{u_s}^r\right)$, $u_{s-1}\le u_s$, as a transformed hidden s-th phoneme ψ_{ks}^{r1} from the k-th word regular transcription $j_{0q_k}=\left(\psi_{k1}^{r1},\psi_{k2}^{r1},\ldots,\psi_{ks}^{r1},\ldots,\psi_{kq_k}^{r1}\right)$. The probability of that that an observed subsequence $\Phi_{s_{k-1}s_k}^{r}=\left(\varphi_{s_{k-1}+1}^r,\varphi_{s_{k-1}+2}^r,\ldots,\varphi_{s_k}^r\right)$, where $(s_k-s_{k-1})=l$ is length of the observation, is a realization of the hidden k-th word transcription $j_{0q_k}=\left(\psi_{k1}^{r1},\psi_{k2}^{r1},\ldots,\psi_{ks}^{r1},\ldots,\psi_{kq_k}^{r1}\right)$ assigns to a product of independent distortions maximized by hidden phoneme ψ_{ks}^{r1} bounds $\{u_s\}$:

$$P\left(\Phi_{s_{k-1}sk}^{r}\big/ j_{0q_k}\right)=\max_{\{u_s\}}\prod_{s=1}^{q_k}P\left(\Phi_{u_{s-1}u_s}^{r}\big/\psi_{ks}^{r1}\right). \qquad (12)$$

In (12) each factor $P(\Phi_{\mu\nu}/\psi)$ is equal to 0 if $\Phi_{\mu\nu}=$ $=(\varphi_{\mu+1}, \varphi_{\mu+2}, \ldots, \varphi_{\nu})$ is not associated with the hidden ψ, otherwise it is computed as a function of both a $\Phi_{\mu\nu}$ to ψ mapping occurrence frequency and acoustic parameter normal laws.

Each Phoneme Recognizer output sequence is processed with the described phonetic-acoustic filter by means of dynamic programming. Therefore, the N>>1 best phoneme sequences of the first level are converted to $N$2>>1 word sequences. The phonetic-acoustic filter parameters are estimated by training samples like in [5].

4. Continuous Speech Interpreter

The Word Recognizer level result is $N$2>>1 best observed phoneme sequences $\Psi_{0\hat{Q}^r}^{r1}=\left(\psi_1^{r1},\psi_2^{r1},\ldots,\psi_v^{r1},\ldots,\psi_{\hat{Q}^r}^{r1}\right)$, r1=1:$N$1, $N$1>>1 and associated with them word sequences $J_{0R}^{r2}=\left(j_1^{r2},j_2^{r2},\ldots,j_v^{r2},\ldots,j_R^{r2}\right)$, r2=1:$N$2, $N$2>>1. On the Speech Interpreter level syntax, semantics and pragmatics are taken into account and among $N$2>>1 word sequences the best one is selected and its understanding is performed.

At this level basically are used the linguistic knowledge. Spoken natural language is specified by WordNet [6] or by means of semantic network for Slavic languages [1].

The simplest method is the following [1]. All conceivable sentences can be packed into subject fields. In turn, all sentences of each subject field (SF) are divided into categories on the basis of transmitted meaning. Each subject field is corresponded with quite a little number of meaning categories.

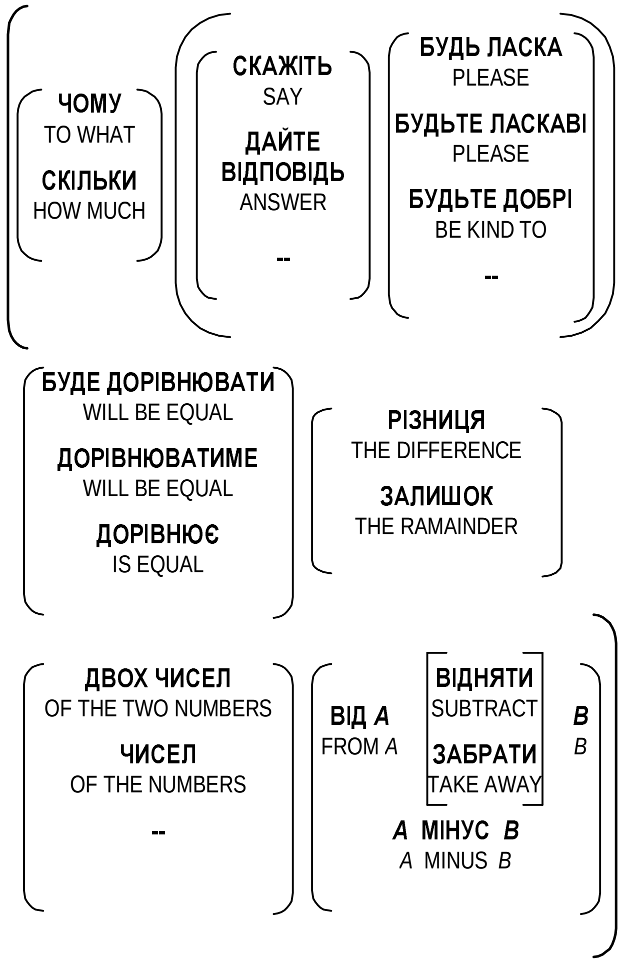
The following meaning categories may apply to the information desk of an airport: questions related to flight arrival; questions related to flight departure; questions related to seat availability; questions related to itinerary; questions related to the location of services at the airport, etc.

Each meaning category (MC) consists of its own set of sentence types. The sentence type (ST) is the construction that economically specifies a set of sentences, which are obtained from one sentence by independent substitutions and inversions for separate words or wordage. A basic element of a sentence type is the subdictionary. They are named accordingly to the SF semantics.

Each MC has quite a little number of sentence types. It is apparent that every MC might be, if necessary, filled out with new sentence types.

All sentence types are easy to specify using list structure languages like *LISP*.

Here is an example of a sentence type for a question related to the difference of two numbers for Ukrainian:



The parentheses ( ) contain invertible subdictionaries, while square brackets [ ] contain non-invertible subdictionaries. Subdictionaries can be inverted only within "superior" parentheses ( ). The sentence type, as a rule, is parametric. In this example, the parameters are numbers, which are the operands A and B. The symbol "--" denotes an empty word.

It is easy to see that, even if we do not to take into consideration variable numbers for the operands A and B, the given ST specifies totally 6!·(2·(2·(3·4))·3·2·3·3) = = 1 866 240 different sentences which are permissible in Spoken Ukrainian and express the same meaning about the difference of two numbers. Among them there are such sentences:

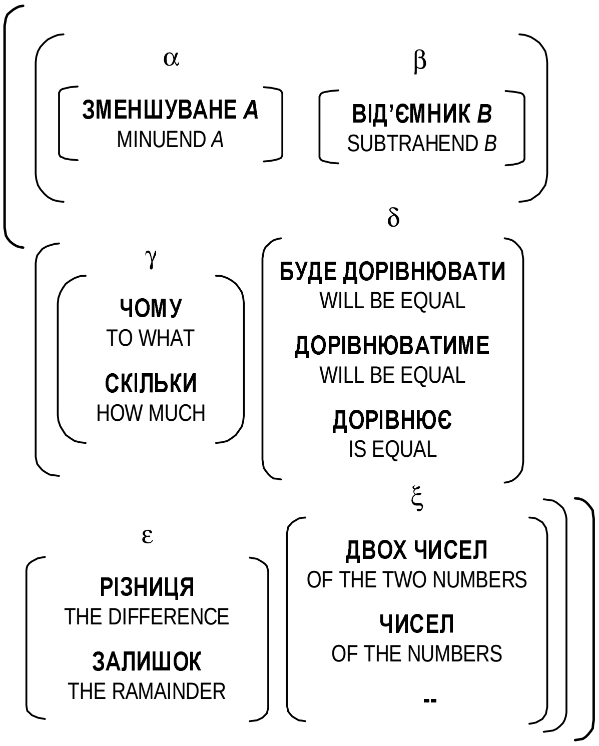
ЧОМУ СКАЖІТЬ ДОРІВНЮЄ РІЗНИЦЯ ЧИСЕЛ *A* МІНУС *B*,

ЧОМУ ДОРІВНЮВАТИМЕ РІЗНИЦЯ *A* МІНУС *B* СКАЖІТЬ БУДЬТЕ ЛАСКАВІ,

СКІЛЬКИ ВІД *A* ВІДНЯТИ *B* БУДЕ ДОРІВНЮВАТИ ЗАЛИШОК ДВОХ ЧИСЕЛ,

ВІД *A* ВІДНЯТИ *B* ЧОМУ ДОРІВНЮЄ ЗАЛИШОК ДАЙТЕ ВІДПОВІДЬ.

Below is given the second example of ST for the MC about the difference of two numbers:



For this ST the subdictionary nominating is performed, namely: α is a minuend, β is a subtrahend, γ is a question word, δ is an action, ε is an operation, and ξ is an action object.

Some examples of generated sentences by the above ST:

ЗМЕНШУВАНЕ *A* ВІД'ЄМНИК *B* ЧОМУ ДОРІВНЮЄ РІЗНИЦЯ ЧИСЕЛ

РІЗНИЦЯ ДВОХ ЧИСЕЛ ДОРІВНЮВАТИМЕ ЧОМУ ВІД'ЄМНИК *B* ЗМЕНШУВАНЕ *A*

Meaning categories and sentence types will be used in the multi-level multi-decision continuous speech understanding process. Here it is emphasized that ST structures are convenient to generate words, which continue permissible initial word subsequences.

While processing each of *N*2 sentences is tested to a ST relation. If no relation detected the sentence is rejected from the further consideration. Otherwise, the relevant MC is assigned and is appended to the list of understanding result pretenders. The result of automatic speech recognition and understanding is that word sequence together with respective ST and MC that has the best probability value among J_{0R}^{r2}, *r*2=1:*N*2, *N*2>>1.

5. Experiments

Two experiments were performed to simulate the three-level ASR system. In the first experiment, only one decision at the first level and multiple decisions for higher two levels were considered.

Firstly a speaker voice file (passport) [4] was formed and a conventional HTK-based automatic phoneme recognition was carried out [7]. The alphabet contains 55 basic Ukrainian phonemes including a phoneme-pause. A speaker pronounced the phonetically rich training sample of above 2113 words containing 20353 phoneme realizations in each of three microphones having unlike acoustic features. Acoustic models accordingly to Section 2 were trained and refined for each basic phoneme, particularly taking into account its both acoustic variability and occurrence. Each phoneme model had three states and 1 to 6 Gaussian mixtures.

The phoneme recognizer output firstly was used to estimate acoustic-phonetic parameters for the second level accordingly to Section 3. Grapheme-to-phoneme converter was used on the basis of two million orthographical words.

For the third level, 15 subject areas and about 3600 sentence types were constructed proceeding from phrasebook sources. City and street names, medicine titles etc. (in respective cases) were considered as variables for the sentence types.

Results have shown that 100% phrase interpretation is not attainable even for multiple decisions on 2nd and 3rd levels. Then N>>1 decisions for the first level were introduced. This and other improvements were implemented.

Final experiment results, particularly recognition and understanding score dependence on N, $N2$, $N3$ are coming up.

6. Conclusion

More adequate acoustic model for speech recognition is a phoneme-threephone model since the co-articulation factor is considered. The phoneme-threephone model operates with $|\Phi|^3$ generative grammars and calculation grows up to $|\Phi|^2$ times comparing to the monophone model, besides, processing a phoneme-threephone grammar that is not free takes additional computations. Therefore, it is expedient to choose N up to $|\Phi|$ and even more to attain comparable memory and computation expenses.

The prospective models looks also phoneme-diphone, syllable- or morpheme-based models [2],[8] supplemented with multiple decisions and this is actual for highly inflected languages with relatively free word order and Slavic languages are among them.

The problem remains of how to guaranty that the optimal solution is not lost in multiple decisions.

Thus, the problem of selecting a speech pattern on the 1st level of the proposed model (phonemes, diphones, syllables, morphemes etc.) is a subject for our further research as well as speech patterns on 2nd and 3rd levels (stress and intonation groups, simple and compound sentences, sentence types, subject areas etc.). As a possible way it is admitted unification for the 2nd and 3rd levels when the lexical-semantic processor filters the improper decisions out.

The 1st level output array looks like an extremely informative object to be explored.

Particular attention should be paid to the 1st level output array carrying extremely useful information about possible phoneme sequences.

7. References

[1]. T.K. Vintsiuk, *Analysis, Recognition and Understanding of Speech Signals*, Kiev: Naukova Dumka, 1987, 264 p (in Russian).

[2]. Taras K. Vintsiuk, “Two Approaches to Create a Dictation/Translation Machine”, *Proceedings of the 2nd International Workshop “Speech and Computer”, SPECOM’97*, Cluj-Napoca, 1997, pp 1–6.

[3]. Taras K. Vintsiuk, Generative Phoneme-Threephone Model for ASR, *Proceedings of the 4th International Conference “Text, Speech and Dialogue”, TSD’2001*, Zelezna Ruda, 2001, pp 201–207.

[4]. Taras K. Vintsiuk, Mykola M. Sazhok, Speaker Voice Passport for a Spoken Dialogue System, *Proceedings of the 3rd International Workshop "Speech and Computer", SPECOM’98*, St.-Petersburg, 1998, pp 175–178.

[5]. Mykola Sazhok, Generative for Decoding a Phoneme Recognizer Output, *Accepted to the Proceedings of the 8th International Conference “Text, Speech and Dialogue”, TSD’2005*, Karlovy Vary, 2005.

[6]. http://wordnet.princeton.edu.

[7]. Young S.J. et al., *HTK Book, version 3.1*, Cambridge University, 2002.

[8]. Bernard Mérialdo, “Multilevel decoding for Very-Large-Size-Dictionary speech recognition”, *IBM J. for R&D, Natural Language and Computing*, 32(2):227–237 (1988).