УДК 004.67

PHONEME RECOGNITION OUTPUT POST-PROCESSING FOR WORD SEQUENCES DECODING

M. Sazhok

International Research/Training Center for Information Technology and Systems

mykola@uasoiro.org.ua

У статті розвивається метод багаторівневого автоматичного розпізнавання мовленнєвого сигналу, що початково був запропонований для флективних мов з вільним порядком слідування слів. Розглянуто два рівні: на першому рівні застосовується пофонемне розпізнавання, результат якого піддається пост-процесингу на другому рівні. Запропонована модель враховує акустичні і фонетичні ознаки, лексикон і особливості спонтанного мовлення. Описується спосіб оцінювання параметрів пост-процесора.

Ключові слова: пофонемне розпізнавання, пост-процесинг результату розпізнавання, флективні мови, спонтанне мовлення, перетворення фонема—графема.

The paper presents advances in a multi-level automatic speech understanding approach that is initially developed for highly inflective languages with relatively free word order. Two levels are considered. On the first level it is applied a phoneme recognizer, which output is post-processed at the second level. The proposed model of post-processing involves acoustic and phonetic features together with lexicon and spontaneous speech peculiarities. A way to estimate the post-processor parameters is described.

Keywords: phoneme recognition, recognizer output post-processing, highly inflective languages, spontaneous speech, phoneme-to-grapheme conversion.

В статье развивается метод многоуровневого автоматического распознавания речевого сигнала, который первоначально был предложен для флективных языков со свободным порядком следования слов. Рассмотрены два уровня: на первом уровне применяется пофонемное распознавание, результат которого подвергается пост-процессингу на втором уровне. Предложенная модель учитывает акустические и фонетические признаки, лексикон и особенности спонтанной речи. Описывается способ оценивания параметров пост-процессора.

Ключевые слова: фонемное распознавание, пост-процессинг результата распознавания, флективные языки, спонтанная речь, преобразование фонема—графема .

Introduction. In accordance to the multi-level multi-decision speech understanding system structure discussed in [1] an approach when continuous speech is firstly recognized as a phoneme sequence and then this phoneme sequence is recognized and understood as a word sequence and meaning appears constructive.

Despite some criticism of this approach, since the best method of speech signal understanding consists in its simultaneous recognizing and understanding, constructing such a multi-level system is a real possibility to distribute the research job between experts in acoustics, phonetics, linguistics and informatics. Moreover, the phoneme recognition must not be rigid but controlled in such a way to yield the best result of understanding.

Apparently, the multi-level speech understanding structure looks as if particularly corresponding for advancing in speech-to-text conversion for a series of highly inflected languages with relatively free word order, and Slavic ones are among them.

If the model retains applicability for languages with other statistical characteristics it means that this approach can be taken to create common implementation of ASR for a wide set of languages in combination with the approach targeting to remove language dependency in speech processing.

In previous research it were analyzed several modifications of the basic structure and described advances in phoneme recognition [2, 3]. In this work the focus is on phoneme recognition result processing with the purpose to extract the pronounced words.

In Section 1 we describe the general structure for a three-level multi-decision speech understanding system. In Section 2 some formalization for the post-processor is given. Section 3 is dedicated to the parameter estimation proceeding from the examples.

1. General structure

The general structure of the considered multi-level multi-decision speech recognition technique is shown in Figure 1. It is consists of three parts. These are Generalized Phoneme Recognizer, Recognizer Post-Processor and Continuous Speech Interpreter.

Generalized Phoneme Recognizer produces N>>1 best recognition responses under free phoneme order or conditioned with constrains on phonemic or morphemic level [3]. Then Phoneme Recognizer Post-Processor analyzes 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.

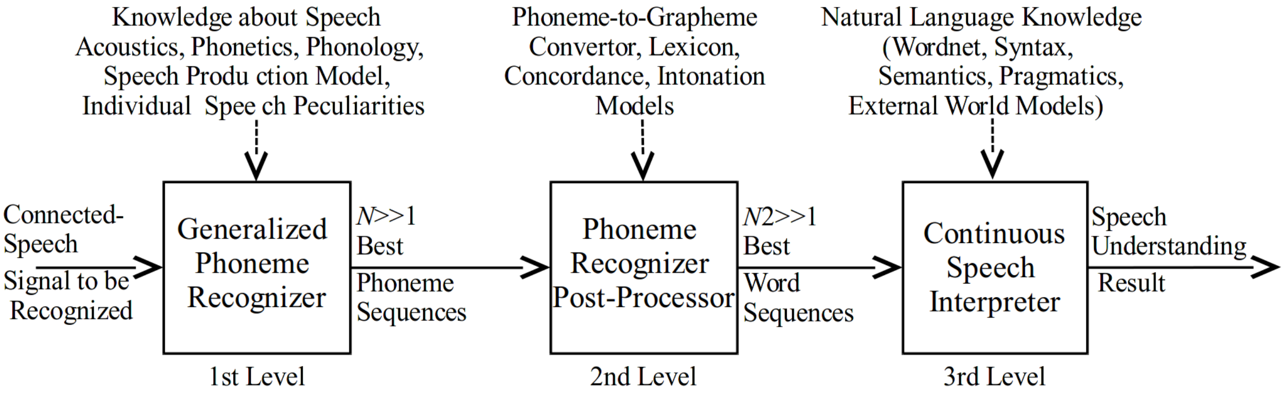
Knowledge about Speech Acoustics, Phonetics, Phonology, Speech Produ ction Model, Individual Spee ch Peculiarities
Phoneme-to-Grapheme Convertor, Lexicon, Concordance, Intonation Models
Natural Language Knowledge (Wordnet, Syntax, Semantics, Pragmatics, External World Models)
Connected-Speech Signal to be Recognized
Generalized Phoneme Recognizer
N>>1 Best Phoneme Sequences
Phoneme Recognizer Post-Processor
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

To focus on Second Level, which tries to extract hypothetical word sequences by phoneme sequences, we must find correspondences between phoneme sequences observed by Phoneme Recognition Post-Processor and the pronounced words composed of, actually, hidden phonemes that could be either misrecognized by First Level or mispronounced by a speaker or distorted by the acoustic channel or noise.

2. Post-processing modeling

One of the Generalized Phoneme Recognizer level result is N>>1 best observed phoneme sequences $\Phi^{r}_{0Q^r} = \left(\varphi^r_1, \varphi^r_2, \ldots, \varphi^r_u, \ldots, \varphi^r_{Q^r}\right)$, r=1:N where Q^r is length of the r-th observed sequence. Moreover, as the result of the First Level, each phoneme observation φ^r_u might be accomplished with information about its duration d^r_u, probability ΔF^r_u and may be other parameters like energy, pitch movement etc. Therefore, actually, as the output of First Level, we consider sequences of certain phonetic events rather than phoneme sequences.

At Second Level we construct a post-processor that must extract, for all $\Phi^{r}_{0Q^r}$, r=1:N, total $N$1>>1 hidden phoneme sequences $\Psi^{r1}_{0Q^{r1}} = \left(\psi^{r1}_1, \psi^{r1}_2, \ldots, \psi^{r1}_s, \ldots, \psi^{r1}_{Q^{r1}}\right)$, r1=1:$N$1, $\psi \in \Psi \equiv \Phi$ and correspond them to word sequences $J^{r2}_{0Q^{r2}} = \left(j^{r2}_1, j^{r2}_2, \ldots, j^{r2}_k, \ldots, j^{r2}_{Q^{r2}}\right)$, $r2$ = 1:$N$2, $N$2>>1 and $j^{r2}_k \in J$ where J is a word dictionary. To avoid loosing the actual word sequence, $N$2>>1 recognition responses are taken.

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

$$P\left(\Phi^r_{s_{k-1}sk} \big/ j_{0q_k}\right) = \max_{\{u_s\}} \prod_{s=1}^{q_k} P\left(\Phi^r_{u_{s-1}u_s} \big/ \psi^{r1}_{ks}\right). \qquad (1)$$

Here each factor expressed as $P(\Phi_{\mu\nu}/\psi)$ is equal to 0 if $\Phi_{\mu\nu} = (\varphi_{\mu+1}, \varphi_{\mu+2}, \ldots, \varphi_{\nu})$ does not correspond to 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 sequence of phonetic-acoustic events is processed with the introduced filter by means of dynamic programming as it is shown in Figure 2. Here we observe 4 phonemes (phonemic events), φ_1, φ_2, φ_3 and φ_4, produced by Generalized Phoneme Recognizer. The observed phonemes can be generated by one or more transcriptions that consist of hidden phones ψ_1, ψ_2, ψ_3 and ψ_4 where ψ_1 generates one phoneme, ψ_2 generates a sequence of two phonemes, ψ_3 generates a sequence of three phonemes and ψ_4 generates an "empty" phoneme, which models phoneme skipping. Solid lines show deterministic transitions between hypothetically generated phoneme events. Dotted lines denote permissible transitions between hidden phonemes. These transitions are conditioned by pronunciation dictionary, word sequence and spontaneous speech peculiarities models [4].

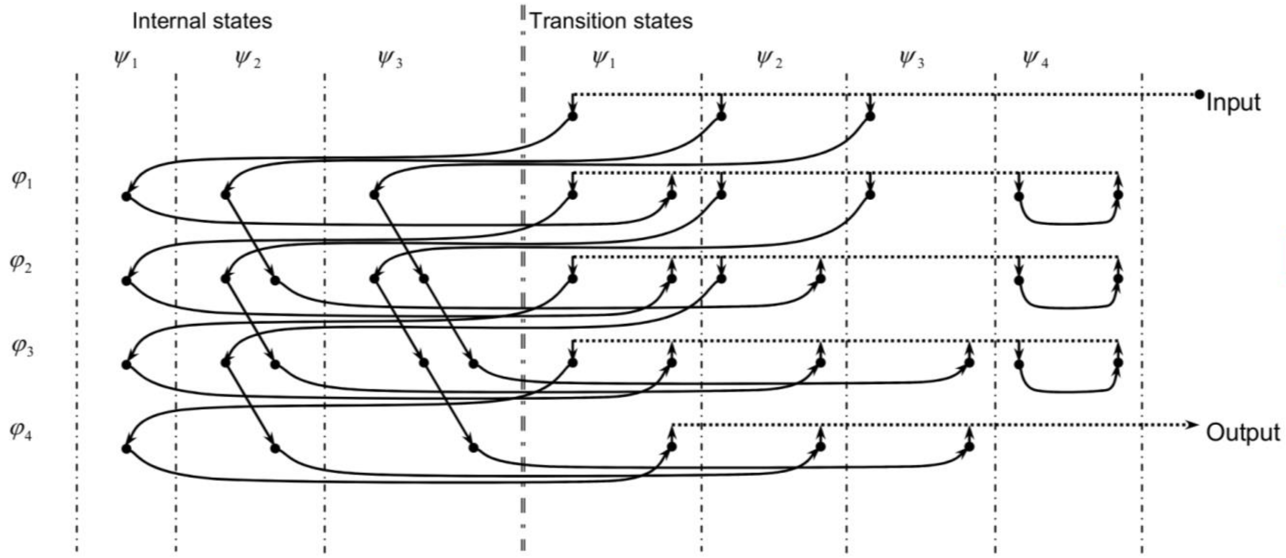

Figure 2: Graph example for the phoneme recognition output post-processor generative model.

An example of how transitions between hidden phonemes are controlled is shown in Figure 3 where speech signal segment corresponding to the pronounced Ukrainian word поверне́те (you will turn) has been converted to the phoneme sequence p o v a r n e1 t y1.

In the example we construct a directed graph each node of which contributes to hypothetical word beginnings based on extracted hidden phoneme sequences. Thus, the second recognized phoneme o corresponds to hidden o, o1, e or y, which potentially belong to words starting with по, пе and пи according to phoneme-to-grapheme conversion [5].

A node corresponding to the last symbol in the valid word contains a complete word marked with (=) that prohibits or (_) that admits word extension to left or right. Mark (!) designates symbol sides where neither valid word may begin or end. To write out a consecutive prospective word we concatenate symbol segments with coinciding marks (!) or (_) until a valid word composed. Dimmed nodes designates no further valid transition exists.

As follows from Figure 5, three word sequences can be extracted: повар не ти, повернете and по варна ти. These hypotheses, alongside with integral criteria, are to be passed to Third Level where the final decision is made considering syntax and semantic knowledge.

Note that using the described technique we may model spontaneous speech feature like word breaks or false-starts in accordance to ideas described in [4] without vocabulary inflation. For doing this we introduce eventual starts of new words even after the incomplete word written out.

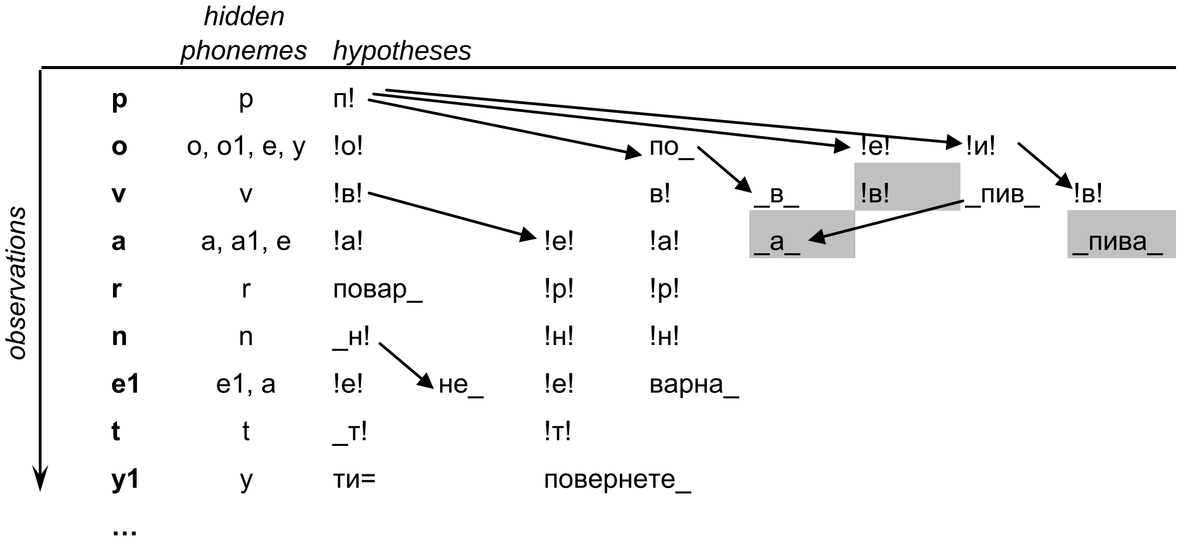

Figure 3: Example of permissible transitions between hidden phonemes controlled with a dictionary.

3. Parameter estimation

As it follows from (1) we actually consider two sub-levels: phonemic and lexical. Such decomposition can be interpreted so that phonemic sub-level is responsible for hypothetical hidden phoneme sequence generation with no reference to the vocabulary. Further, once a hypothetical phoneme has been appended to a partial sequence, the lexical sub-level checks the updated partial phoneme sequence for validity referring to hypothetical (partial) word sequences.

To estimate phonemic sub-level parameters we use a training set obtained by the phoneme recognizer for sentences with known text.

In Figure 4 we show an example of the phonemic sub-level model structure. Here we admit that a hidden phoneme ψ_t can be observed in form:

$$M_{\psi} = \left(\psi^{-}, \psi, {}^{+}\psi\right), \ \psi \in \Psi \cup \varnothing \tag{2}$$

By means of this model a hidden phoneme is able to generate:

- a phoneme with the name may not match the hidden phoneme name;

- sequences of two phonemes both of which may not match the hidden phoneme;
- sequences of three phonemes where the central phoneme matches the hidden phoneme;
- an empty phoneme (*), which means that the hidden phoneme is skipped.

To weaken the proposed constrains (2), we may consider generated sequences of three phonemes where neither phoneme name matches the hidden phoneme name and longer sequences may be considered as well.

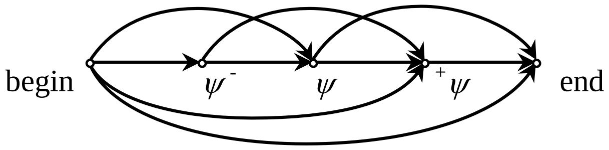

Figure 4. Generalized structure for phonemic sub-level model estimation.

Each model needs to be initialized with phoneme names and to be updated with acoustic features like length and criteria. We propose to initialize the models in a way that allows for reducing the total amount of models and keeping all prospective phoneme sequences.

Table 1 illustrates an example of parameter estimation for hidden phones by the sample of Ukrainian sentence: ми не глухі (we aren't deaf). The reference phoneme sequence is: *pau* m y1 n e h l u kh1 i1 sp, where pau and sp means silence that is always hidden at sentence boundaries. The example is based on real phoneme recognition results. For each hidden phoneme we construct one or more models. Thus, proceeding from the 4th to 7th observed phonemes we construct models for hidden phonemes n and e that generate n and n+y1 for n and y1-e and e for e.

Table 1

Phonetic level model hypotheses initialization

Observation				Model	
No	Length	Criteria	Name	Name	Hypotheses
1	51	-22.15	pau	pau	pau
2	16	-30.19	m	m	m
3	9	-34.10	y1	y1	y1
4	8	-30.01	n	n	n, n+y1
5	3	-33.17	y1		
6	4	-32.71	e	e	y1-e, e
7	8	-28.27	h	h	h, h+l
8	6	-31.31	l	l	l, l+l
9	7	-35.32	l	u	l, *
10	21	-27.98	kh1	kh1	kh1, l-kh1, kh1+i, l-kh1+i
11	11	-29.78	i		
12	18	-26.95	i1	i1	i1, l-i1
13	15	-24.31	t1	sp	sp

Already initialized models just update their statistics for length, criterion and occurrence. Factor $P(\Phi_{\mu\nu}/\psi)$ from (1) is estimated proportionally to the occurrence of phoneme sequence $\Phi_{\mu\nu}$ generated by hidden phoneme ψ. After all models (2) have been initialized by the training sample, the expectation-maximization iterations [6] are applied until the criterion (1) converges or another stop condition reached.

Conclusion. The proposed technique allows for finding the regularities by which hidden phoneme sequences transform to observed phoneme event sequences. It is shown that, beside highly inflexed languages, spontaneous speech recognition may benefit from multi-level multi-decision approach application. The deficiency of each post-processor is its activation after the end of the basic process. So the ways to integrate the post-processor scheme in the computation of nodes of phoneme recognition graph should be considered.

References

1. 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.

2. N. Vasylieva, M. Sazhok T. Vintsiuk, G. Chollet. Acoustic-Phonetic Model Application for Syllable Speech Recognition Output Post-Processing. Proceedings of the 12th International Conference SpeCom’2007, Moscow, 2007, pp. 182-187.

3. Н. Васильєва. Використання граматик вільного порядку слідування фонем і складів для пофонемного розпізнавання злитого мовлення. Штучний інтелект. – Донецьк, 2011, № 4, с. 80-86.

4. Робейко В.В. Моделирование особенностей спонтанной украинской речи в системах автоматического речевого сигнала. // Кибернетика и вычислительная техника: Межведомственный сборник научных трудов. – Вып. 170. – Київ, 2012. – С. 76—85.

5. V. Robeiko, M. Sazhok. Bidirectional Text-To-Pronunciation Conversion with Word Stress Prediction for Ukrainian. In Proc. UkrObraz’2012, Kyiv, 2012, pp. 43-46.

6. T.K. Vintsiuk. Analysis, Recognition and Understanding of Speech Signals. - Kiev: Naukova Dumka, 1987, 264 P. (in Russian)