Three-Level Multi-Decision Model for 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 multiple decisions, but under simplified conditions, at all levels of a speech signal processing hierarchy is 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 phoneme 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's 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 phoneme sequence recognition responses. Then a Generalized Word Recognizer analyses these phoneme sequences in order to generate *N*2>>1 possible word

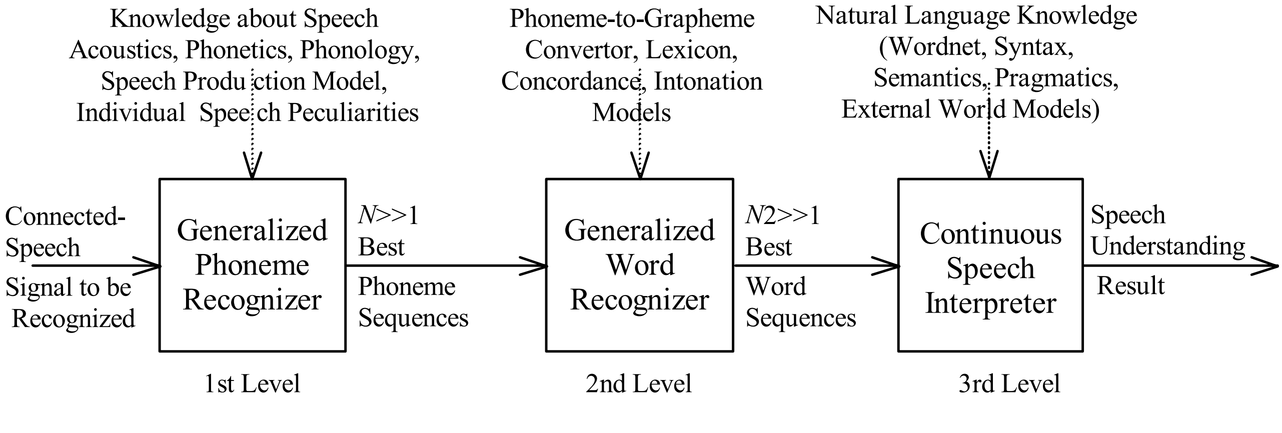

Knowledge about Speech Acoustics, Phonetics, Phonology, Speech Production Model, Individual Speech Peculiarities
Phoneme-to-Grapheme Convertor, Lexicon, Concordance, Intonation Models
Natural Language Knowledge (Wordnet, Syntax, Semantics, Pragmatics, External World Models)
Connected-Speech Signal to 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

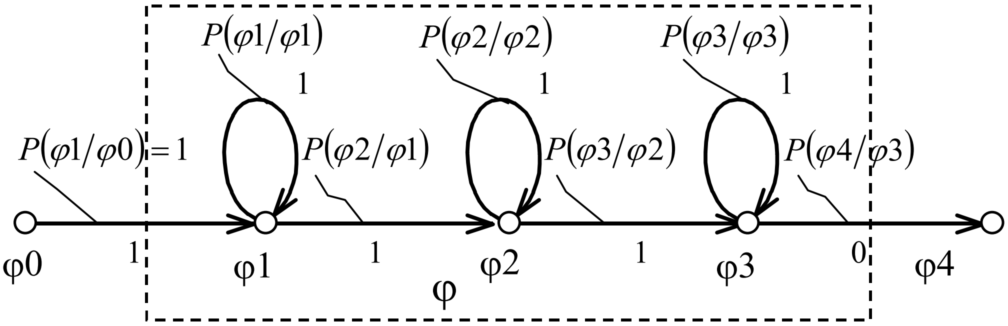
Figure 1. Three-Level Speech Recognition System Structure.

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.



$$P(\varphi1/\varphi1)+P(\varphi2/\varphi1)=1 \qquad P(\varphi2/\varphi2)+P(\varphi3/\varphi2)=1 \qquad P(\varphi3/\varphi3)+P(\varphi4/\varphi3)=1$$

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

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

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 $P(\varphi1/\varphi1)$, $P(\varphi2/\varphi2)$, $P(\varphi3/\varphi3)$. 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].

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:

$$\begin{aligned} P(X_{\mu\nu}/\varphi) = &\max_{\{(w_1,w_2)\}} \\ &\left\{\left[(P(\varphi1/\varphi1))^{w_1-\mu-1}(1-P(\varphi1/\varphi1))\prod_{i=\mu+1}^{w_1}P(x_i/\varphi1)\right]\times\right. \\ &\times\left[(P(\varphi2/\varphi2))^{w_2-w_1-1}(1-P(\varphi2/\varphi2))\prod_{i=w_1+1}^{w_1}P(x_i/\varphi2)\right]\times \\ &\left.\times\left[(P(\varphi3/\varphi3))^{\nu-w_2-1}(1-P(\varphi3/\varphi3))\prod_{i=w_2+1}^{\nu}P(x_i/\varphi2)\right]\right\} \end{aligned} \tag{1}$$

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.

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

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

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

To perform generalized phoneme recognition means to find N>>1 best phoneme sequences. The phoneme sequence generalized algorithm based on the criterion (2) is described in [5]. The result of the generalized phoneme recognizer is N>>1 best extracted 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 a length of the r-th extracted sequence.

Under the recognized phoneme sequences there are hidden phoneme and respective word sequences which a speaker mentioned to pronounce. The way to reveal these possible sequences is the purpose of procedures introduced on the next levels.

3. Generalized Word Recognizer

The Phoneme Recognizer level produces on its output N>>1 best 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 a length of the r-th sequence, which are observations for the Generalised Word recognizer. 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 estimations of parameters like energy, pitch movement etc. So each each observed on the output of the 1st level we consider as a phonetic-acoustic event, which together hide phoneme and respective word sequences to be extracted and subsequently interpreted.

At the second level Word Recognizer must extract for all $\Phi^r_{0Q^r}$, r=1:N total N1>>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:$N1$, $\psi \in \Psi \equiv \Phi$ and associate them with 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:$N2$, $N2$>>1 and $j^{r2}_k \in J$ where J is a word dictionary. To avoid loosing the actual word sequence we take $N2$>>1 second level recognition responses.

Thus, we interpret observed phoneme 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 $\Phi^r_{s_{k-1}s_k} = \left(\varphi^r_{s_{k-1}+1}, \varphi^r_{s_{k-1}+2}, \ldots, \varphi^r_{s_k}\right)$, 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 a product of independent distortions maximized by hidden phoneme ψ^{r1}_{ks} bounds $\{u_s\}$:

$$P\left(\Phi^r_{s_{k-1}sk} \middle/ j_{0q_k}\right) = \max_{\{u_s\}} \prod_{s=1}^{q_k} P\left(\Phi^r_{u_{s-1}u_s} \middle/ \psi^{r1}_{ks}\right). \qquad (3)$$

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 in accordance to the describer criterion by means of dynamic programming. The extraction procedure therefore contains two components one of which is responsible for generation of permissible phoneme sequence transformation (acoustic-phonetic filter) and the other one that provides lexical knowledge.

The phonetic-acoustic filter parameters are estimated by training samples like in [5]. The lexical part should provide some kind of conversion between pronunciation and spelling and contain a dictionary or a word building model.

Thus, N>>1 best phoneme observation sequences of the first level are converted to $N2$>>1 word sequences.

Having several hypothetical word sequences we intend to choose the one most appropriate according to semantics, syntax and pragmatics. And this is a job for the third, final, level for continuous speech interpretation.

4. Continuous Speech Interpreter

The Word Recognizer level result is $N2$>>1 best phoneme sequences $\Psi^{r1}_{0\hat{Q}^r} = \left(\psi^{r1}_1, \psi^{r1}_2, \ldots, \psi^{r1}_v, \ldots, \psi^{r1}_{\hat{Q}^r}\right)$, $r1$=1:$N1$, $N1$>>1 and associated with them word sequences $J^{r2}_{0R} = \left(j^{r2}_1, j^{r2}_2, \ldots, j^{r2}_v, \ldots, j^{r2}_R\right)$, $r2$=1:$N2$, $N2$>>1, which are observations for the Continuous Speech Interpreter level. On this level syntax, semantics and pragmatics are taken into account and among $N2$>>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 [7] 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 a subdictionary, which is 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 proposed to specify using list structure languages like *LISP* [5].

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 $N2$ 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}, $r2$=1:$N2$, $N2$>>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 the conventional HTK-based automatic phoneme recognition was carried out [8]. 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. Depending on model pruning strategy we extracted from 1 to 5 thousand models that makes tens of model per phoneme in average.

Ukrainian spelling-pronunciation bidirectional converter was used on the basis of two million orthographical words. The converter is based on the *n*-gram mapping derived from pronunciation rules.

The other phoneme recognizer output (isolated words and phrases) was used as a control sample for the second level. The generalized word recognizer has two main modifications concerning a usage of the lexical component: simultaneously or consequently. In the first case hundreds solutions are needed to have a right one among them, and in the second case tens of best solutions were sufficient.

6. Conclusion

More adequate acoustic model for speech recognition is a phoneme-triphone model since the co-articulation factor is considered. The phoneme-triphone model operates with $|\Phi|^3$ generative grammars and calculation grows up to $|\Phi|^2$ times comparing to the monophone model, besides, processing a phoneme-triphone 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 sub-word prospective models looks also phoneme-diphone, syllable-, morpheme-based or language independent data-driven models [9] supplemented with multiple decisions and this is actual for multilingual ASR and, particularly, 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.

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]. Taras K. Vintsiuk, Mykola M. Sazhok, Multi-level Multi-decision Models for ASR, *Proceedings of the 10nd International Workshop "Speech and Computer", SPECOM’2006*, Patras, Greece, 2006, pp. 69–76.

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

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

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

[9]. G. Chollet, K. McTait, D. Petrovska-Delacrétaz. Data Driven Models to Speech and Language Processing. In:

Lecture Notes in Computer Science, Vol. 3445, 2005,
pp. 164-198.