Generalized Automatic Phonetic Transcribing of Speech Signals

Taras K. Vintsiuk

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

ABSTRACT

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

АБСТРАКТ

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

INTRODUCTION

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

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

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

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

PHONEME RECOGNITION IN CONTINUOUS SPEECH. GENERAL IDEA

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

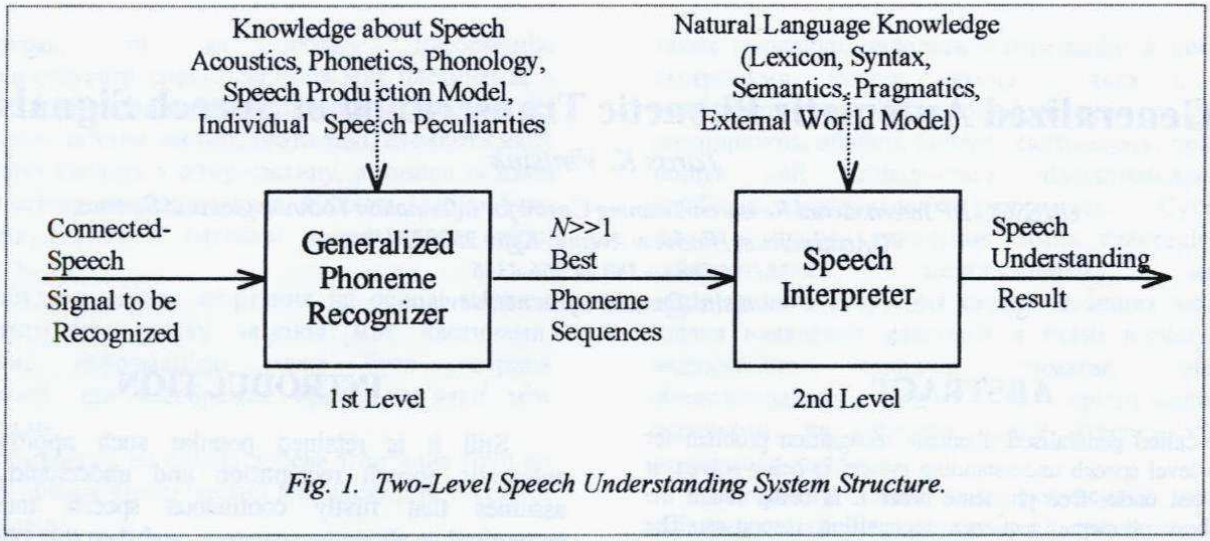

Fig. 1. Two-Level Speech Understanding System Structure.

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

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

GENERAL FREE PHONEME-THREEPHONE SEQUENCES GENERATIVE GRAMMAR

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

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

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

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

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

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

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

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

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

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

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

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

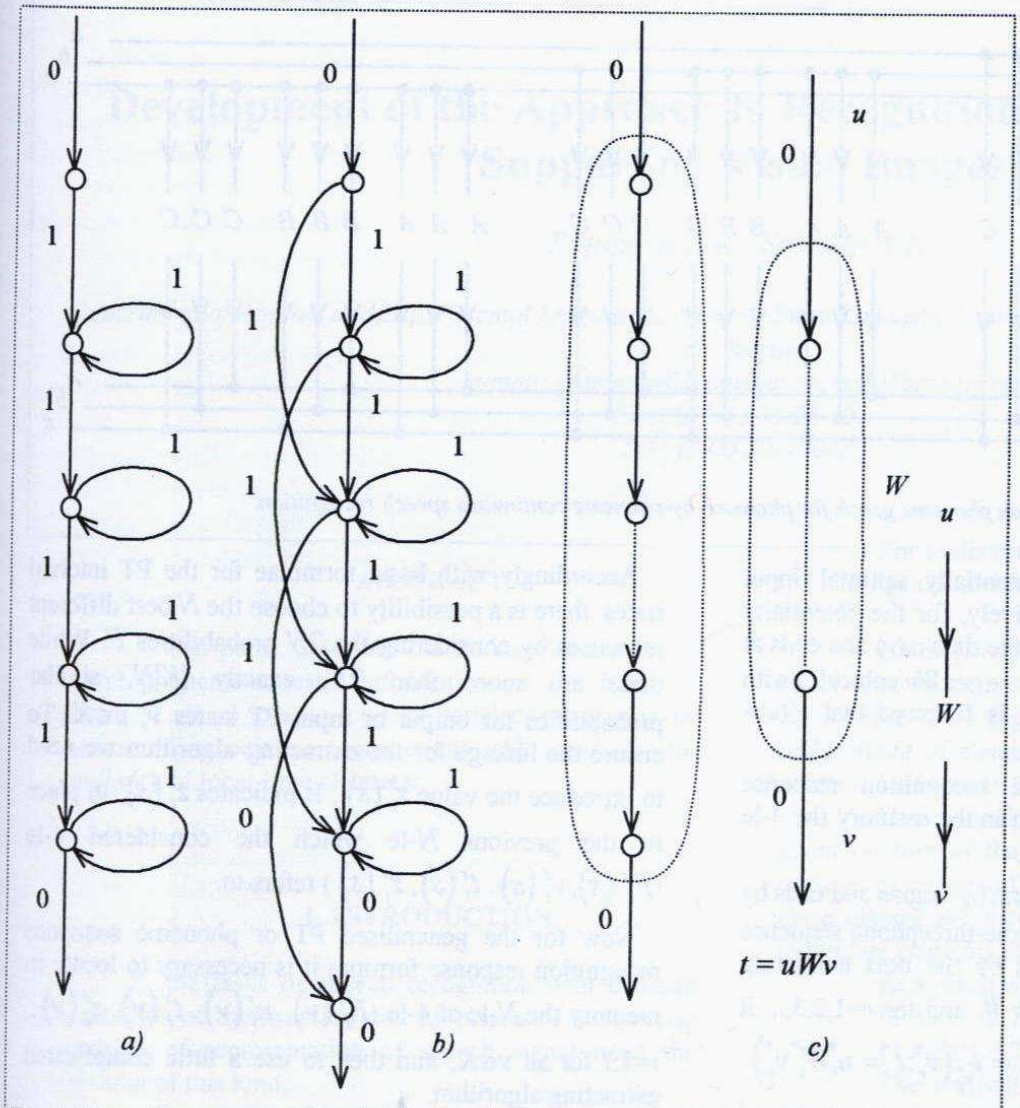

Fig. 2. Generative grammars (graphs) for the phoneme-threephone: a) no microelement omis-sion; b) no two microelements running omission; 3) schematic notes of the PT graph t = uWv.

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

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

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

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

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

PHONEME SEQUENCE RECOGNITION ALGORITHM

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

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

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

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

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

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

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

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

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

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

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

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

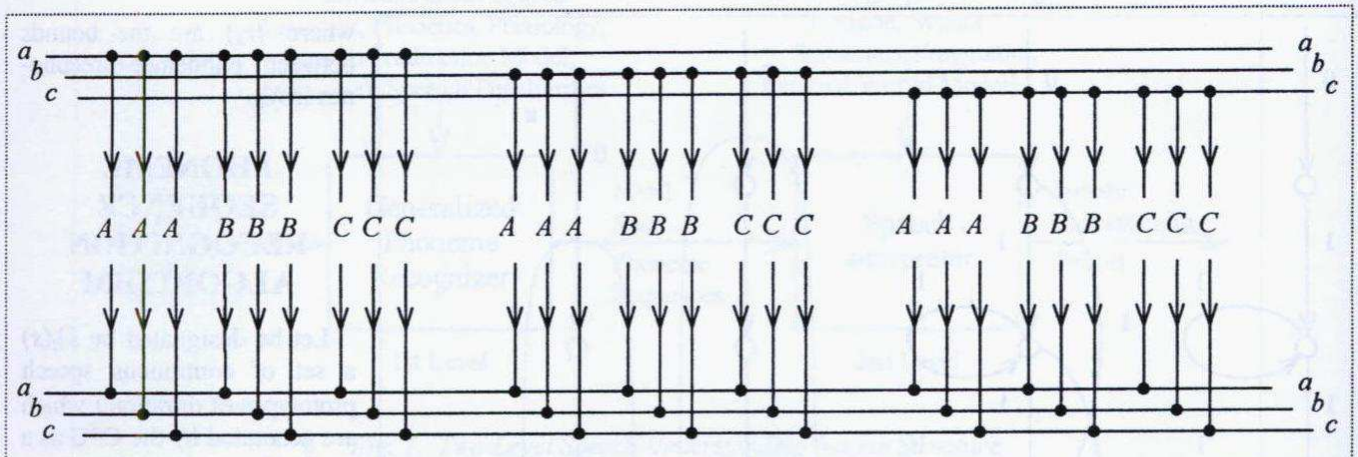

Fig. 3. Common phoneme graph for phoneme-by-phoneme continuous speech recognition.

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

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

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

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

THE GENERALISED ALGORITHM

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

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

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

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

CONCLUSION

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

ЛІТЕРАТУРА

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