PHONEME-BY-PHONEME RECOGNITION AND SEMANTIC INTERPRETATION OF MULTI-SPEAKER SPEECH (THE HCDP-APPROACH)

TARAS VINTSYUK

Speech Recognition and Synthesis Laboratory
Institute of Cybernetics
Kiev, Ukraine, USSR 252207

ABSTRACT

A new approach to phoneme-by-phoneme recognition and semantic interpretation of multi-speaker speech is proposed. The approach is based on a constructive (C) representation of complex speech signals with hierarchic (H) structure of speech patterns (signals, microphonemes, phonemes, diphones, syllables, words, sentences, communicated senses). The recognition and semantic interpretation reside in composing (C) for a given speech signal and subsequent parsing of such complex speech pattern that is consistent with all levels of the hierarchy and is the most similar in some sense to the one to be recognized. The guided composition and subsequent parsing of this complex speech signal are realized by means of dynamic programming (DP). Some examples of solved problems are listed.

SPEECH PATTERN HIERARCHY AND MATHEMATICAL MODELS OF SEGMENTS

The HCDP-method integrates the approved principles of speech information decoding and processing and generalizes the CDP-method /1/, /2/, /3/. The hierarchic principle presumes the hierarchy of the patterns. The speech signals are discribed by sequences of observable elements-vectors x_i: $X_1 = (x_1, x_2, \ldots, x_i, \ldots, x_1)$, where 1 is a length of the speech signal in uniform or quasi-uniform discrete time with spacing (mean spacing) of 15 ms for instance. The subsequences of the elements $X_{mn} = (x_{m+1}, x_{m+2}, \ldots, x_n)$ being named segments are interpreted as the speech patterns or more precisely as the realisations of first-level speech patterns (the microphonemes, phonemes, diphones, or syllables). In this case X_{mn} is considered as the first-level segment. Sets of the signals X_{mn} for the first-level patterns j^1 are specified by distributions $p(X_{mn}/j^1)$, $j^1 \in J^1$, where J^1 is an alphabet of the first-level patterns. The second-level speech patterns j^2 are specified by the transcriptions in the alphabet of the first-level patterns: $j^2 = (j^1_1, j^1_2, \ldots, j^1_s, \ldots, j^1_{q(j^2)})$, where $q(j^2)$ is the length of the transcription of the pattern $j^2 \in J^2$, J^2 is the alphabet of the second-level patterns. The second-level segments correspond to the second-level patterns and are composed (the composition (C)) of the first-level segments by merging them into the sequences in conformity with the second-level pattern transcription. For instance if the microphonemes or phonemes are the first-level patterns then the phonemes or diphones (syllables) can be the second-level patterns correspondingly.

The patterns and segments at the next hierarchic levels (the syllables, words, sentences, communicated senses) are defined similarly. Let $j^r \in J^r$, $j^r = (j^{r-1}_1, j^{r-1}_2, \ldots, j^{r-1}_s, \ldots, j^{r-1}_{q(j^r)})$, $p(X_{mn}/j^r)$ be the r-level pattern from the alphabet J^r, the transcription of the pattern j^r and the probability of the r-level segment X_{mn} under

the condition of the pattern j^r correspondingly. The top-level patterns in the hierarchy - the communicated sense from a given finite set of senses - are specified by a canonical form and a formal constraction being named a directed semantic network and sense types and sentence types /2/, /3/. While forwarding to publications /2/, /3/ for the details let us concentrate attention on a fact that the top-level hierarchic patterns - the communicated sense - are specified actually by a list of the sentences that express the same sense. But this specification is realized by some memory-saving means instead of direct enumeration. From this more accurate definition it also follows that the r-level pattern is not obligatory expressed with one transcription in the alphabet of the (r-1)-level patterns and there can be several or even many such transcriptions.

Constructive (C1) nature of the model manifests in expressing the probabilities of the segments X_{mn} under the condition of the r-th pattern $p(X_{mn}/j^r)$ as products of the probabilities of the corresponded to the transcription j^r segments under the condition of the patterns of the previous (r-1)-th level:

$$p(X_{mn}/j^r) = \prod_{s=1}^{q(j^r)} p(X_{m_{s-1}m_s}/j_s^{r-1}) \quad , \quad (1)$$

where m_s are bounds of the (r-1)-level segments: $m_o=m$, $m_{s-1}<m_s$, $m_{q(j^r)} = n$. Thus the probability of the observed (to be recognized) signal $X_1 = X_{o1}$ under the condition of the top-level hierarchic pattern $j^R \in J^R$, J^R is the alphabet of the top-level hierarchic patterns, takes a form of the product of the corresponding segment probabilities under the condition of the first-level patterns:

$$p(X_{o1}/j^R) = \prod_{s=1}^{\bar{q}(j^R)} p(X_{m_{s-1}m_s}/j_s^1) \quad . \quad (2)$$

In the expression (2) $\bar{q}(j^R)$ is a number of the first-level patterns from the sequence of which the top-level hierarchic pattern j^R is composed, $m_o = 0$, $m_{s-1}<m_s$, $m_{\bar{q}(j^R)} = l$ are the bounds of the first-level segments.

To describe (specify) the mathematical model of the speech signals and to use it then for solving the speech recognition problems there is obviously sufficient to give the transcriptions of the patterns at all levels of the hierarchy with the segment distributions under the condition of all first-level hierarchic patterns $p(X_{mn}/j^1)$, $j^1 \in J^1$. These distributions are specified for every possible segment length n-m that takes generally different values for the different patterns $j^1 \in J^1$.

In line with the expression (1) it may seem that the segments of the speech signal are considered as mutually independent ones. In reality it follows just from the expression (1) as well as (2) that there is a strong deterministic dependence of the segments in the sequences that is manifested in constraints on the pattern order in the sequences, i.e. is expressed in the transcriptions of the patterns at all levels of the hierarchy.

RECOGNITION CRITERION AND METHOD

By using the maximal likelihood method let us classify the signal X_1 to be recognize as such top-level hierarchic speech pattern that the acceptable for the subject field sequence of the first-level patterns that is composed by the transcriptions in accordance with the pattern hierarchy will induce on X_1 such first-level segmentation for which the likelihood expression reaches an absolute maximum:

$$j^R(X_1) = \operatorname*{argmax}_{j^R \in J^R} \max_{\{m_s\}} \prod_{s=1}^{\bar{q}(j^R)} p(X_{m_{s-1}m_s}/j_s^1). \quad (3)$$

The expression (3) presumes the exhaustive search through all pattern transcriptions if the patterns are specified not by one but by two or more transcriptions.

The recognition criterion (3) determines top-down and down-top analysis of the signal X_1 simultaneously. It is important that by solving the problem (3) one receives a consistent with all hierarchic levels interpretation referring if necessary to the segment borders of all-level hierarchic patterns being contained in the analizing signal.

By analizing the expression (3) and taking account of a fact that the borders of the r-level segments unconditionally coincide with the borders of certain (r-1)-level segments one concludes that the exhaustive search to maximize the expression (3) can be avoided and the solution can be found by the Bellman's optimality principle with help of the dynamic programming. For computation it is more convenient to use a logarithm of the likelihood. The expression (3) is transformed into

$$j^R(X_1) = \operatorname*{argmax}_{j^R \in J^R} \max_{\{m_s\}} \sum_{s=1}^{\bar{q}(j^R)} \ln p(X_{m_{s-1}m_s}/j_s^1). \quad (4)$$

The constructivity (C2) of the HCDP-method is just in referencing the effective method to maximize (4) for the segment borders $\{m_s\}$ and all-level hierarchic patterns - in using the dynamic programming (DP) for these goals.

To afford the constructivity C2 one needs the constructive (C3) techniques to specify the hierarchy of the patterns and their transcriptions and the constructive (C4) means to describe the distributions $p(X_{mn}/j^1)$, $j^1 \in J^1$ under the condition of the first-level patterns for every possible segment length. Let us consider the realization of the constructivity principles with the particular examples from /2/, /3/.

MICROPHONEMIC RECOGNITION AND SEMANTIC INTERPRETATION

The first level of the hierarchy is the microphonemes (parts of the phonemes). The microphoneme j^1 is specified by one or more standard elements being denoted by $e(j^1)$ and having more frequently the same physical nature as the observed speech elements. The distribution of the segment X_{mn} under the condition of the microphoneme j^1 is defined by the relationship:

$$G(X_{mn}, j^1) = \ln p(X_{mn}/j^1) = \sum_{i=m+1}^{n} \ln p(x_i/e(j^1)) = \sum_{i=m+1}^{n} g(x_i, e(j^1)), \quad (5)$$

where the segment length satisfies the condition

$$T_{min}(j^1) \leqslant n-m \leqslant T_{max}(j^1) \quad . \quad (6)$$

In accordance with (5)-(6) one considers the quantity $g(x_i, e(j^1))$ as an elementary measure of similarity between the observed element x_i and standard element $e(j^1)$, and $G(X_{mn}, j^1)$ as the similarity between the segment X_{mn} and the first-level pattern j^1 such that the latter itself is the stationary segment being composed of one element $e(j^1)$ that is replicated n-m times to quote the constraints (6). The number of the microphonemes $j^1 \in J^1$ is 128, 256, 512, but not greater than 1024.

The second level of the hierarchy is the words that are specified by one or more so-called acoustic or Q-transcriptions - the sequences that are composed of the first-level patterns /2/, /3/.

The third level of the hierarchy is the arbitrary word sequences being composed of the free-ordered words from a selected vocabulary. The fourth level is the allowable sentences of the subject field that are specified by the sentence types, or sence types, or directed semantic network /2/, /3/. The fifth level is a canonic form of the communicated sense.

By restricting to the first two or three levels a system is obtained to recognize correspondingly the words or continuous speech that is composed of the words from the chosen vocabulary.

PHONEME-BY-PHONEME (DIPHONIC) RECOGNITION AND SEMANTIC INTERPRETATION

The diphonic model of speech signal generation /2/, /3/ is a good compromise reflecting dynamic properties of the speech signals and realizing the phonemeness principle in the recognition. Let us insert in the hierarchic model being dealt in the previous section an additional level - the level of the diphones that takes an intermediate place between the level of the microphonemes and the level of the words. The diphonic word transcriptions are evidently defined by their phonetic transcriptions in a unique manner. The obtained six-level speech recognition and semantic interpretation system is realized the phoneme-by-phoneme recognition principle more evidently.

ZERO LEVEL OF THE HIERARCHY - MULTIDIMENSIONAL (VECTOR) QUANTIZATION

The constructivity (C5) of the HCDP-method is in using the principle of the vector quantization of the speech signals, i.e. in inserting the zero-level hierarchic patterns where the observed sequences $X_1=(x_1, x_2,\dots,x_i,\dots,x_1)$ from the vectors-elements x_i are replaced by the sequences $I_1=(j^o_1, j^o_2,\dots,j^o_i,\dots,j^o_1)$ from vectors-scalars $j^o_i=j^o(x_i)$: each observed element-vector x_i is replaced by a number of a domain $j^o_i = j^o(x_i)$ to which the observed element x_i belongs in the multidimensional space of the signals x, $j^o \in J^o$, where J^o is the alphabet of the zero-level patterns.
The introduction of the zero-level patterns allows to go over from an investigation of the relationships in the vector sequences to the investigation of the relationships in the sequences of the scalars. Now one ought to substitute the distributions $p(I_{mn}/j^1)$, $j^1 \in J^1$, where $I_{mn}=(j^o_{m+1}, j^o_{m+2},\dots,j^o_n)$ for the distributions $p(X_{mn}/j^1)$ in the formulas (1)-(5). Then in line with the principle C4 one should point out the constructive principles of specifying the distributions $p(I_{mn}/j^1)$ for the allowable values of n-m. The first group is the methods based on a tabular specification of the distributions $p(I_{mn}/j^1)$, on an effective storing these distributions in the networks, or simply on storing the encountered values I_{mn} under the condition of the pattern $j^1 \in J^1$. In the second group there are the methods based on an approximation of the distributions $p(I_{mn}/j^1)$ with help of simple expressions and on usage of the formulas that are analogous with (5). One example:

$$p(I_{mn}/j^1)=\prod_{i=m+1}^{n} p(j^o_i/j^1) \text{ or } G(I_{mn},j^1)=$$
$$=\sum_{i=m+1}^{n} \ln p(j^o_i/j^1).$$

Here the distributions $p(I_{mn}/j^1)$ are specified by the tables of $|J^o| * |J^1|$ numbers $p(j^o/j^1)$.

LEARNING TO RECOGNITION AND MULTI-SPEAKERNESS

The necessary knowledge base - such a priori data as the pattern hierarchy, subject field, syntax, semantics, vocabulary, alphabets and transcriptions of the upper-level patterns - is prepared beforehand by a creator of the speech signal recognition systems. The remained undefined data (the alphabets of the lower-level patterns, the corresponding transcriptions of the lower-level patterns, the distributions $p(X_{mn}/j^1)$ or $p(I_{mn}/j^1)$ for all first-level hierarchic patterns) are computed in a learning-to-recognition mode from a multi-speaker learning set.

References

/1/ T.K.Vintsiuk, CPD-methodes de reconnaissance et d'interpretation de la parole, "Le Symposium Sovietico-Francais sur "Le Dialogue Acoustique de l'Homme avec la Machine", Moscou, 1984, p. 38 - 41.
/2/ T.K.Vintsiuk, Speech recognition and semantic interpretation, "Kibernetika", 1982, No.5, p. 101 - 111 (in Russian).
/3/ T.K.Vintsiuk, Analysis, recognition and interpretation of speech signals, Kiev, "Naukova Dumka", 1987, 280 p. (in Russian).