Generative Model for Decoding a Phoneme Recognizer Output

Mykola Sazhok

Int. Research/Training Center for IT and Systems, Kyjiv 03680 Ukraine,
mykola@uasoiro.org.ua

Abstract. The paper presents a way to advance to a multi-level automatic speech understanding system implementation. Two levels are considered. On the first level a free (or relatively free) grammar phoneme recognition is applied and at the second level an output of the phonemic recognizer is automatically interpreted in a reasonable way. A Generative Model approach based model for phoneme recognizer output decoding is proposed. An experimental system is described.

1 Introduction

In accordance to the multi-level 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 (Fig. 1) 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.

Apparently, the multi-level speech understanding structure looks as if particularly corresponding for advancing a creation of dictation machines and spoken dialog systems for a series of highly inflected with relatively free word order languages, and Slavic ones are among them.

Obviously, the output of the Phoneme Recognizer level must imply a potential of its further processing. It means that a phoneme sequence produced by recognizer must be readable in sense of machine. This does not mean that it must be human-readable but the latter ability is by all means prominent. Besides, a machine, unlike a human, might intensively use other parameters the recognizer extracted from speech like phoneme length, amplitude etc.

Thus, the problem of the next by Phoneme Recognizer levels is to learn a machine to interpret an acquired phoneme sequence or to find a hidden phoneme sequence that is an actual transcription of the pronounced utterance. In terms of Generative Model [2], we must suggest a way to generate all possible phoneme sequences associated with the proper permissible sequences and to compare them with observation. This is exactly what is investigated in next two sections where appropriate models are justified and a training procedure is described.

How to attain the required phoneme recognition results and whether appropriate models and algorithms are available nowadays? Such a system is attainable due to system parameters refinement and speaker individuality modeling by means of Speaker Voice Passport and we describe it in the experimental section.

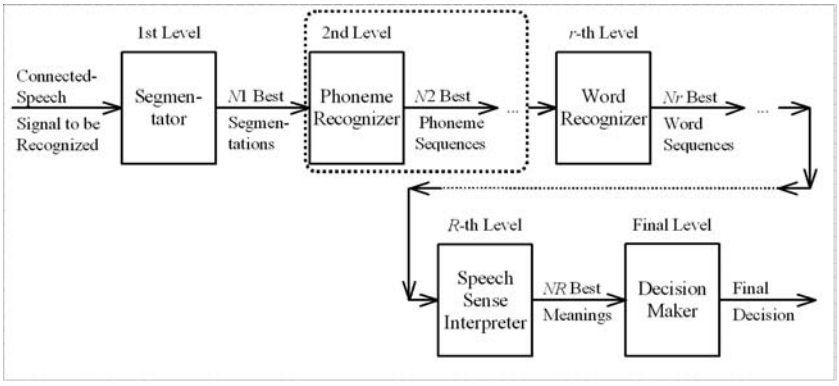

Fig. 1. Multilevel multi-decision Dictation/Translation Machine structure

2 Association model between a phoneme sequence pronounced and the one generated by the phoneme recognizer

We consider a recognizer output as a sequence of phoneme observations. Each phoneme observation associates phoneme name ö, duration d, energy E, likelihood g and may be more acoustic parameters the recognizer might extract from speech like pitch etc. We specify the machine readability of the recognizer output or the phoneme observation sequence as a possibility to generate its model by a phoneme transcription obtained from the pronounced text. Phoneme sequences associated with both observation and its generated model must match.

Let us consider an operator transforming a generated automatically by text phoneme sequence $(\varphi_1, \varphi_2, ..., \varphi_q)\, \varphi_k \in \Phi, 0 \le k \le q$ to all permissible sequence of phoneme observations of a given length l:

$$v^l(\varphi_1, \varphi_2, ..., \varphi_q) = \left(v^{l_1}(\varphi_1), v^{l_2}(\varphi_2), ..., v^{l_q}(\varphi_q)\right), \sum_{k=1}^{q} l_k = l, 0 \le l_k \le \overline{l}, \quad (1)$$

where $\overline{l}$ value means an upper length of the phoneme sequence replacing the phoneme φ_k and each $v^{l_k}(\varphi_k)$, $0 \le k \le q$, generates a subsequence of phoneme observations by a given phoneme φ_1 with length l_k :

$$v^{l_k}(\varphi_k) = \left(w_1^{l_k}(\varphi_k), w_2^{l_k}(\varphi_k), ..., w_{l_k}^{l_k}(\varphi_k) \right), \tag{2}$$

where model phoneme observation $w_s^{l_k}(\varphi_k)$, $0 \le s \le l_k$ follows from the associated with φ_k model, which structure will be considered later.

Equating l_k to 0 means the phoneme φ_k is substituted with zero-length phoneme sequence or dropped. To eliminate 2 running omissions the following restriction must be satisfied:

$$\overline{\exists} k, \;\; 0 \le k \le q-1 : \;\; l_k = 0 \; and \; l_{k+1} = 0. \tag{3}$$

To force a phoneme subsequence associated with $v^{l_k}(\varphi_k)$ taking out certain length $\underline{l} \le \overline{l}$ to contain the phoneme φ_k we require existence of one and only one $w_S^{l_k}(\varphi_k)$ associated with φ_k, $0 \le S \le l_k$:

$$\forall l_k, \; \underline{l} \le l_k \le \overline{l} \;\; \exists ! S, \; 0 \le S \le l_k : \varphi_k \prec w_S^{l_k}(\varphi_k), \tag{4}$$

where $\prec$ is an association sign and state S is the terminal state of a model.

Thus, we introduce a model of the phoneme φ_k observation as an l_k-state generative model, $0 \le l_k \le \overline{l}$. Each state s, $0 \le s \le l_k$, corresponds to a set of possible phonemes $\psi_{s,i}^k \in \Psi_s^k \in \Phi$ from the sequence with length l_k that substitutes φ_k. Hence, all sequences the k-th model generates may be interpreted as Descarte's product of sets Ψ_s^k by $s = \overline{1, \; l_k}$.

Proceeding from (4), $\Psi_S^k = \{\varphi_k\}$ for terminal state S of the k-th model.
Auxiliary input and output states are introduced to specify permissible transitions between states of adjacent models.
Acoustic parameters are described by their normal law distributions:

$(A, \; \Sigma) = \left(\left(a^d, \sigma^d\right), \; \left(a^E, \; \sigma^E\right), \; \left(a^g, \; \sigma^g\right) \right)$,

where the distributions are specified for phoneme duration d, energy E and likelihood g.

The algorithm of hidden phoneme sequences decoding from speech output is similar to continuous speech recognition with grammar. It appears that a non-iterative algorithm to train the model might be derived.

3 Phoneme observation model training

A proposed algorithm to learn phoneme observation model consists in extracting model prototypes from the found best trajectories on graph and updating final models by their prototypes. Configuration of links between nodes on the graph follows from (1)–(3).

Initially we assume each model prototype has a maximal number of states $\overline{l}$. Proceeding from (4) a terminal state S must be included and one of simplifications proposed is to assume S to a fixed state of the initial model prototype.

To catch a substitution of the phoneme with an empty phone sequence we insert an empty phoneme $\emptyset$ between phoneme observations and choose one of non-terminal state that is applicable to the empty phoneme.

In each graph node we compute an elementary likelihood that is a positive value when model and observation phoneme names coincide at the terminal state and zero otherwise. Therefore, from the human point of view the integral likelihood is proportional to number of common phonemes in the model transcription and in the phoneme recognizer output.

Note that an observation attained from phonemic recognizer is actually divided into two streams: phonetic and acoustic. As far considering initial prototypes we operate only with phoneme names a likelihood for acoustic stream is not available. So we just do collect acoustic data in prototypes. When updating final models by their prototypes the collected acoustic data is used to estimate their acoustic parameters distribution.

Analyzing a graph and likelihood one may conclude that normally multiple trajectories may have the best score (Fig. 2). It means that there exist a k-th model having N_k>1 prototypes in context of one training sample and we keep all this prototypes assigning to them a probability equal to $1/N_k$. This value is accumulated in the respective model as well. After passing all training samples models are to be purged and merged to form a final set of models.

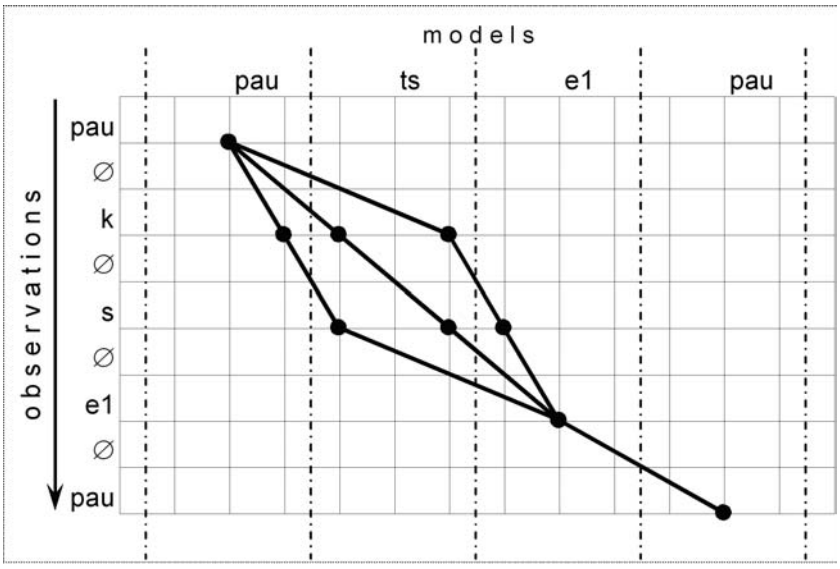

Fig. 2. Graph for phoneme observation models in a training sample. The recognizer output phoneme sequence is 'pau k s e1 pau', under conditions of pronounced word of 'pau ts e1 pau'. The best trajectories are shown. Following the trajectories the model prototypes are extracted. Note, acoustic data of observations is stored to build the global model.

According to the graph illustration given in Fig. 2 we extract model prototypes with the following phonemic descriptions:
1:(PAU, k / pau, 4), 2:(PAU, k / pau, 5), 3:(pau, PAU / pau, 6); 4:(k / 1, ts, 7), 5:(k, s / 2, ts, 8), 6:(s / 3, ts, 9); 7:(s, E1 / 3, e1, 9), 8:(E1 / 4|5, e1, 9); 9:(PAU / 6|7|8, pau).

Here in brackets before the slash a phoneme sequence replacing a model phoneme is indicated. Each phoneme from the sequence is associated with the model state and a capitalized phoneme is associated with the terminal state. From the right of slash a model phoneme name and adjacent model prototypes instances, if applicable, are denoted. Additionally, a probability to each model prototype is assigned.

4 Experimental Training Setup

The experiment was divided into stages of (1) training and control sample preparation, (2) speaker voice file (passport) forming, (3) attaining phoneme recognition output and (4) performing the train procedure for the phoneme observation sequence decoder.

The text of training samples was formed from isolated words extracted from a dictionary of rated Ukrainian words taking into account each phoneme occurrence and acoustic variability. This work is based mainly on [3]. Thus, the training sample text contained 2113 words and total 16127 phonemes except a phoneme-pause. The alphabet contained 55 basic Ukrainian phonemes including both stressed and non-stressed versions of vowels, palatalized versions for all but two consonants and a phoneme-pause. Occurrences of each non-pause phoneme in the training text lied between 10 (palatalized ‘sh’ and ‘zh’) and 1001 non-stressed ‘o’. No short pause model was provided as far the training sample includes only isolated words.

The control sample represented mostly top rated words and less phoneme variability. We just scanned a rate dictionary from the top and took words containing new triphones.

A speaker pronounced the entire training sample in each of three microphones having unlike acoustic characteristics. Acoustic models were trained and refined for each basic phoneme specifically taking into account its both acoustic variability and rate. Each phoneme model had three states and 1 to 6 Gaussian mixtures. So the speaker voice passport was formed.

A free-grammar phoneme recognition procedure was performed on total 11000 words from the control sample. Attained phoneme sequences exposed obvious resemblance with generated automatically by text transcriptions for pronounced words.

Before carrying out the train procedure for the phoneme observation sequence decoder the model parameters were adjusted: $\overline{l}$ and $\underline{l}$ were assigned to 3 and a model prototype terminal state S was fixed at 2 and empty phoneme applicable state is preassigned to 3. Permissible links are specified for each state by

pairs (phoneme, state) indicating start node relatively to the current phone. For instance, (-1, 1) permits link to state 1 of the preceding phone.

Using the developed *Perl* module, all training samples were successfully passed and sets of 3000-5000 models were formed. The module for phoneme recognizer output decoding procedure is under construction.

5 Conclusion

The idea of machine-readable text has been investigated and a model proposed allows for converting phonetic recognizer output to valid phoneme sequences, theoretically, even in case they have no matching phonemes.

We dealt with only one best phoneme sequence of the phoneme recognizer output but actually $N >> 1$ best recognition outputs might be considered. These procedures still take small time amount due to the free phoneme grammar used.

The future plans are to accomplish the global model training and a decoder procedure, to consider multiple decision phoneme recognizer output, to test the approach on continuous speech, to investigate fast speaker voice passport forming and to build models for speech recognition with no vocabulary restrictions.

References

1. Taras Vintsiuk, Multi-Level Multi-Decision Model for Automatic Speech Recognition and Understanding // International Summer School Neural Nets "E.R.Caianiello" 3rd Course "Speech Processing, Recognition and Artificial Neural Networks", 1998, Vietri sul Mare (SA) Italy, pp 341-344.
2. Taras Vintsiuk. Generative Phoneme-Threephone Model for ASR // In: Text, Speech, and Dialogue (4^{th} Int. Conf. TSD 2001, Zelena Ruda, Czech Republic, Proceedings), Lecture Notes in Artificial Intelligence 2166 Subseries of Lecture Notes in Computer Science, Springer, 2001, pp 201-207.
3. Nina Vasylyeva. Training Samples Forming for Automatic Speech Synthesis by Text. - Magister diplom work, Kyjiv 2003, 88 p.