Syllable Speech Recognition Output Post-Processing Based on Models of Acoustics, Phonetics and Lexicon

Mykola Sazhok [1,2], *Nina Vasylieva* [1], *Taras Vintsiuk* [1], *Gerard Chollet* [2]

[1] Department of Speech Science and Technology
International Research/Training Center for Information Technologies and Systems, Kyiv, Ukraine,
[2] Department of Signal and Image Processing (TSI), ENST, Paris, France

mykola@uasoiro.org.ua, ninel@uasoiro.ogr.ua, vintsiuk@uasoiro.org.ua, chollet@tsi.enst.fr

Abstract

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 syllable-based grammar phoneme recognizer, which output is post-processed at the second level. The described model of post-processing involves acoustic and phonetic features together with lexicon. The ways to select a set of sub-word units like syllables and multispeaker corpus used in experimental research are considered. Experimental results, problems and future research are discussed.

Index Terms: syllable recognition, recognizer output post-processing, highly inflective languages, phoneme-to-grapheme.

1. 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 a creation of dictation machines and spoken dialog systems 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 more 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 [2].

In previous research at the 1st level we considered the Generalized Phoneme Recognizer which produced N>>1 best sequences of phonemes accomplished with acoustic estimations under condition of free phoneme order. At the second level, the Generalized Word Recognizer post-processed the output of the previous level. It showed promising experimental results on a single speaker database for isolated word recognition.

The next obvious step to extend the research is introducing a multispeaker database. At the same time we intended to integrate the lexicon in post-processor graph node computing. The latter is considered as a way to reduce number of decisions for the post-processor. The problem of selecting a speech pattern on the level is a subject for this research as well.

In this paper we consider a modification of the three-level multi-decision speech understanding system. The structure of this system is shown in Figure 1. It is consists of three parts. These are Generalized Sub-Word Unit (Syllable) Recognizer, Recognizer Post-Processor and Continuous Speech Interpreter.

The Generalized Syllable Recognizer produces N>>1 best recognition responses under free (or relatively free) syllable order grammar. Then the Syllable Recognizer Post-Processor analyses these phoneme sequences in order to generate $N$2>>1 possible word sequences. By these word sequences a Speech Interpreter makes a decision about the speech understanding response via Natural Language Knowledge.

In Section 2 we justify the selection of sub-word units for the 1st level. In Section 3 formalization for the post-processor is given. In Section 4 we describe the data and knowledge base used for recognition. Section 5 is dedicated to experimental research.

2. Sub-word Unit Selection

In previous works to produce the output of the first level it was used a free grammar phoneme recognizer [1]. Despite the fast implementation, robastness of the latter is far from desirable, specifically for the multispeaker case. Syllables are considered as alternative sub-word units which still weakly depends on a dictionary.

Two ways of syllable selection are analyzed: rule-based and open syllables.

Rule-based syllable selection follows heuristics postulating the placement of syllable boundaries in dependence on coinciding phonemes. Open syllables end with a vowel or a phoneme-pause. Data-driven syllables are in scope of interest as well but not implemented yet.

The syllables have been extracted automatically by the rate dictionary of 137640 words. Although syllable order is free, this is not so for open syllables: syllables ending with a phoneme-pause always follow a syllable ending with a vowel.

Table 1 illustrates that syllable-based grammar significantly improves the phoneme recognition score (up to 1.7 times) comparing to free phoneme order recognition for isolated words. Average length of Ukrainian word is 7,43 phonemes and maximal occurred is 20 phonemes. In all cases recognition done by means of Julius-Julian [3] is performed in real time, which is approximately equal for the considered types of syllables. We should also notice that rule-based syllables look more preferable.

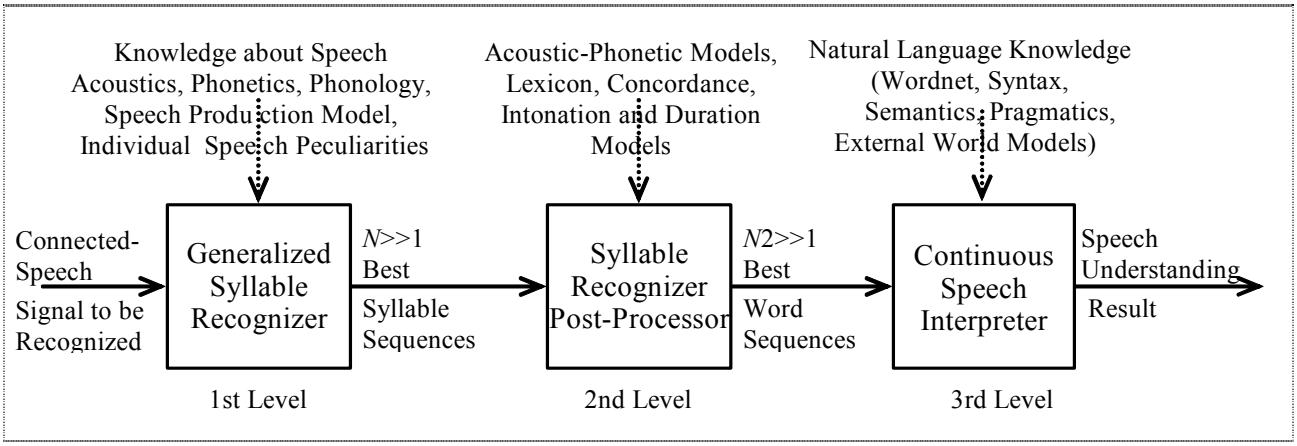

Figure 1: *Graph for phoneme observation models in a training sample*.

Table 1. *Phoneme accuracy in dependence of the sub-word unit type for Ukrainian test samples.*

Test sample	Sub-word unit type	Amount of units	Phoneme accuracy
11000 words	monophone	55	46.0
11000 words	rule-based syllable	9436	79.5
11000 words	open syllable	4966	78.3
100 sentences	monophone	55	49.3
100 sentences	rule-based syllable	9436	56.8
100 sentences	open syllable	4966	55.5

3. Post-processing modeling

The syllable recognizer output is N>>1 best phoneme sequences $\Phi^{r}_{0Q^r}=\left(\varphi^{r}_{1},\varphi^{r}_{2},...,\varphi^{r}_{u},...,\varphi^{r}_{Q^r}\right)$, r=1:N, where Q^r is a length of the r-th observed sequence. Moreover, as the result of the first level, each φ^{r}_{u} is accomplished with estimations of acoustic parameters like duration d^{r}_{u} of the phoneme, its probability ΔF^{r}_{u} and may be other parameters like energy, pitch movement etc. Actually, we consider a sequence of phonetic-acoustic events, which are observed after the syllable recognizer applied.

The aim of the post-processor is to extract for all $\Phi^{r}_{0Q^r}$, r=1:N total $N$1>>1 hidden phoneme sequences $\Psi^{r1}_{0Q^{r1}}=\left(\psi^{r1}_{1},\psi^{r1}_{2},...,\psi^{r1}_{s},...,\psi^{r1}_{Q^{r1}}\right)$, r1=1:$N$1, $\psi\in\Psi\equiv\Phi$ and associate them with word sequences $J^{r2}_{0Q^{r2}}=\left(j^{r2}_{1},j^{r2}_{2},...,j^{r2}_{k},...,j^{r2}_{Q^{r2}}\right)$, r2=1:$N$2, $N$2>>1 and $j^{r2}_{k}\in J$ where J is a lexicon. To avoid loosing the actual word sequence $N$2>>1 recognition responses are taken.

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

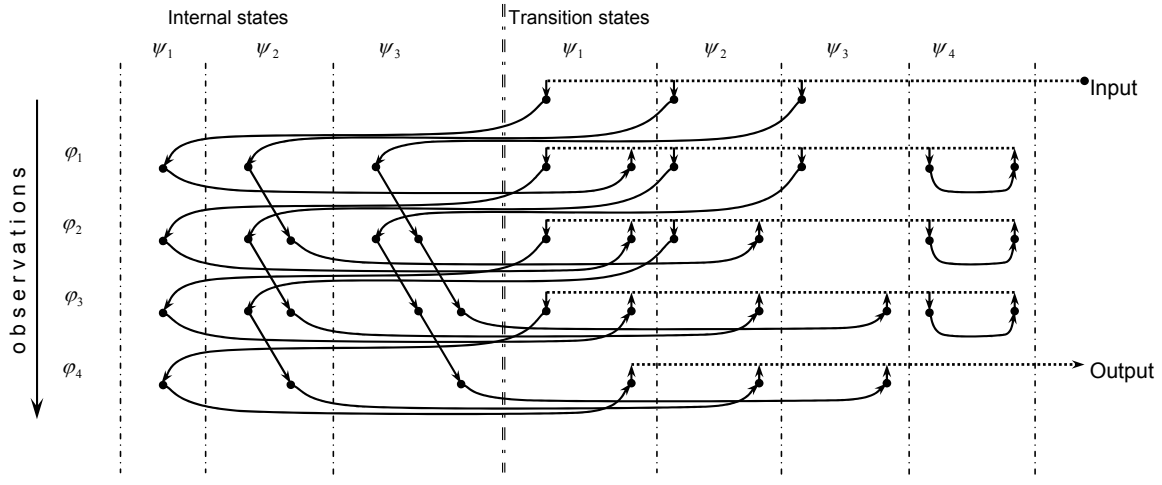

Figure 2: *Graph for the post-processor.*

$$P\left(\Phi^{r}_{s_{k-1}s_k} \Big/ j_{0q_k}\right) = \max_{\{u_s\}} \prod_{s=1}^{q_k} P\left(\Phi^{r}_{u_{s-1}u_s} \Big/ \psi^{r1}_{ks}\right). \quad (1)$$

Here 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 sequence of phonetic-acoustic events is processed with the introduced filter by means of dynamic programming as it is shown in Figure 2. We observe 4 phonemes (phonetic events) produced by the Syllable Recognizer under the condition of three hidden phonemes presented by their acoustic-phonetic models (APMs). The observed phonemes can be replaced with 1 to 3 observed phonemes respectively and a hidden phoneme that is observed as an empty phoneme (omitted). There dotted lines denote inter model transitions provided by the grammar. Solid lines show internal deterministic transitions.

Lexicon part of the post-processor provides constrains on grammar and performs the final phoneme-to-grapheme conversion including word boundaries.

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

The parameters of probabilities (1) that are also APM parameters are estimated by training samples in accordance to [4]. In Fig. 3 we illustrate a graph of APM prototypes extraction for the recognizer output ‘pau k s1 o o pau’, under conditions of pronounced word of ‘pau ts1 o h o1 pau’ (‘of this’ in Ukrainian). The selected trajectories are shown in solid lines and a permissible but not selected trajectory is in dashed line. The following phonemic descriptions are extracted for prototypes:

1:(PAU, k / pau, 4), 2:(PAU, k / pau, 5), 3:(PAU / pau, 6); 4:(k / 1, ts1, 7), 5:(k, s / 2, ts1, 8), 6:(s / 3, ts1, 9); 7:(s, O / 3, e1, 9), 8:(o / 4|5, O, 9); 9:(∅ / 6|7|8, h, 11); 10:(o / 6|7|8, h, 12); 11:(o / 9, o1, 13); 12:(∅ / 10, o1, 13); 13:(PAU / 11|12, pau).

Here in brackets before the slash a phoneme sequenced replacing a model phoneme is indicated. Each phoneme from the sequence is associated with the model state and a capitalized phoneme is associated with a state which associated phoneme matches with the observed one. 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.

Note, acoustic data of observations is updated for each prototype to build the global model iteratively or by purging less probable models.

4. Data and Knowledge Base

Data and knowledge base includes a Ukrainian speech corpus for estimation of parameters for acoustic and acoustic-phonetic models (APMs) and a lexicon as well.

We used the Ukrainian multi-speaker speech corpus that is in stage of its formation. Currently it contains above 30 000 word realizations and thousands of sentences from about 100 speakers living in different regions of Ukraine. The samples keep the phoneme rate proportions and are phonetically balanced [5].

A lexicon contains about 2 millions of word forms that correspond to 151000 of basic forms (lemmas). Actually these lemmas produce above 3 millions of word forms but many of them have same orthography and pronunciation.

On basis of the lexicon and a 250 MB text corpus we generated a word rate dictionary of 157000 words.

The module of phoneme text to orthographic text conversion uses 22 generalized rules of n-gram mapping and refers to the whole lexicon. Actually, on the stage of computing nodes on the graph (Figure 1) this module can be replaced with a phoneme sequences hash (2- and 3-grams) that occurred inside words.

5. Experiments

The experiment was divided into stages of (1) train and control sample preparation, (2) phoneme training, (3) syllable recognition (4) post-processor parameters estimation (5) post-processor testing.

For phoneme training, spoken samples were taken from the Ukrainian multi-speaker speech corpus. We considered isolated words for the phoneme acoustic model parameters estimation. Totally we took about 19858 word samples, and 147 445 phone samples, except a phoneme-pause, from 70 speakers.

The alphabet chosen contains 55 basic 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 30 (palatalized ‘sh’ and ‘zh’) and 1200 for non-stressed ‘o’. No short pause model was

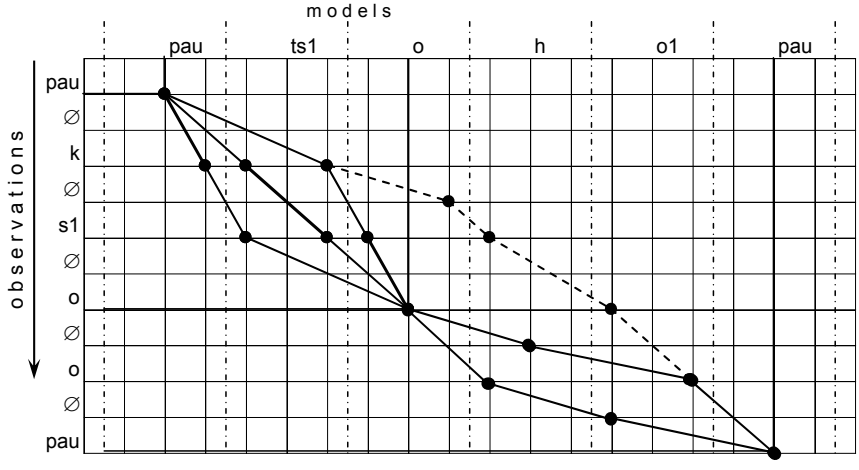

Figure 3: *Graph for phoneme observation models in a training sample.*

provided as far the training sample includes only isolated words.

Acoustic models were trained and refined by means of HTK [6] for each of 55 monophones, particularly taking into account acoustic variability and occurrence of the phoneme. Obtained phoneme models have three states and 4 to 12 Gaussian mixtures.

Syllable recognition was accomplished by means of Julian [3] for two sets of syllables: open-syllables and rule-based syllables with respective grammars in accordance to Section 2.

The output of syllable recognizer including phoneme segmentation and criterion was used to estimate parameters of acoustic-phonetic models (APMs). We used speech data from a single speaker not included in phoneme acoustic models training. The words selected for data is a result of the rate dictionary scanning from the top and taking words containing new triphones. Totally 8000 of such words were recorded and first 3000 are recorded twice. We estimated different amount of models as indicated in Table 2.

Before testing the post-processor we adjusted its lexical parameters: constrains on vowel stress were alleviated and word sequence output was allowed. In all experiments a full vocabulary was used (2 mln. word forms).

Post-processor was tested on different sets of isolated words, though output was allowed as *N*2-best word sequences.

Table 2. *Results of post-processing procedure.*

Syllable type	APM training corpus	Total/used APMs	Test corpus	*N*2	WER %
Rule	5000	3700/3300	6000 words	5	4.7
Open	5000	3900/3300	6000	5	5.2
Rule	5000	3700/3300	6000	7	4.5
Open	5000	3900/3300	6000	7	4.8
Rule	11000	7500/7000	2100*3	5	4.9
Open	11000	7900/3300	2100*3	5	5.1
Rule	11000	3700/3300	100 sentences	5	18.2

It follows that the post-processing accuracy is about 95% for isolated words. Significant degradation on sentences is caused by numerous short words with high criterion. The penalty on inter word transitions might improve the situation.

6. Conclusion

The considered model is actual for highly inflected languages with relatively free word order and Slavic languages are among them.

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 contains even more constrains so additional computations are taken. Therefore, it is expedient to choose N up to $|\Phi|$ and even more to attain comparable memory and computation expenses.

The problem remains of how to guaranty that the optimal solution is not lost in multiple decisions.

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 syllable recognition graph should be considered.

Data-driven syllables or alternative sub-word units are the target of future research.

7. Acknowledgements

This research is carried out in frames of the INTAS Grant, Contract No YS05-109-4212, and Ukrainian multi-speaker speech corpus was formed in frames of the Grant of the President of Ukraine for a Gifted Youth, Contract No 32 from May 30 2006.

8. References

[1] Taras K. Vintsiuk, Mykola M. Sazhok, “Multi-Level Multi-Decision Models in ASR”, Proc. of the 10th International Workshop "Speech and Computer", SPECOM’2005, Patras, 2005, pp. 69–76.

[2] Gérard Chollet, Kevin McTait, Dijana Petrovska-Delacrétaz, Data Driven Approaches to Speech and Language Processing. G. Chollet et al. (Eds.): Nonlinear Speech Modeling, LNAI 3445, pp. 164–198, 2005.

[3] A. Lee, T. Kawahara and K. Shikano, "Julius – an open source real-time large vocabulary recognition engine." In Proc. European Conference on Speech Communication and Technology (EUROSPEECH), pp. 1691–1694, 2001.

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

[5] Nina Vasylyeva, Mykola Sazhok, “Text Selection for Training Procedures under Phoneme Units Variety”, Proc. of the 10th International Workshop "Speech and Computer", SPECOM’2005, Patras, 2005, pp. 629–632.

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