1

Extra Large Vocabulary Continuous Speech Recognition Algorithm based on Information Retrieval

Valeriy Pylypenko
International Research/Training Center for Information Technologies and Systems
Kyiv, Ukraine

1. Introduction

There exists a necessity for speech recognition with a huge numbers of alternatives.
For example, during the voice control of a computer it is impossible to predict the subsequent word on the basis of several previous ones because it is defined by control logic, instead of text properties. From the other hand there is a necessity for growth of the volume of the dictionary to capture all possible synonyms of the same command caused by difficulty for users to remember the single command name variant.
The next example concerns the text dictation. The application of such systems is limited by the texts, which are statistically similar to one where statistics were collected. Additional spoken editing of the text demands the presence of all words in the actual dictionary.
Thus, there are applications where it is desirable to have a dictionary as large as possible, in future to cover all words for the given language (for some languages more than 10M words).
The additional information to restrict the number of alternatives can be received from a speech signal immediately. For this purpose it is proposed to execute preliminary trial recognition by using the phonetic transcriber. Phonemes sequence analysis allows to build the queries flow. Applying the information retrieval approach considerably limits the number of alternatives for recognition.

2. The baseline recognition systems

The approach is applicable for any recognition system where phonemes and phoneme recognition (phonetic transcriber) are present but the number of phonemes no more than approximately 500 units.
As reference systems HMM-based HTK (Young et al., 2006) and Julius (Lee, 2009) toolkits are used.

3. ELVIRS Algorithm for isolated words

3.1 Architecture

The architecture of the system is shown in Figure 1. The *features extraction* and *acoustic models* blocks are reused from the baseline system. Common *pattern matching* unit with subset of

Source: Advances in Speech Recognition, Book edited by: Noam R. Shabtai,
ISBN 978-953-307-097-1, pp. 164, September 2010, Sciyo, Croatia, downloaded from SCIYO.COM

vocabulary is used on the second pass. Changes are concentrated in the new first recognition pass when *phonetic transcriber* is applied to make the sequence of phonemes. Then *information retrieval procedure* builds the sub-vocabulary for the second pass.

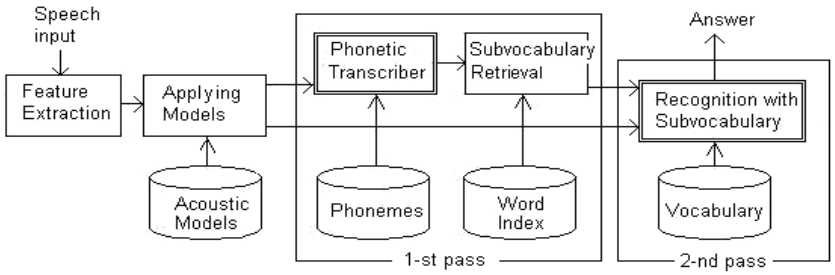

Fig. 1. The architecture of ELVIRS recognition system

3.2 Phoneme recognizer

The phonetic transcribing algorithm (Vintsiuk, 2000; Vintsiuk, 2001) builds a phonetic sequence for speech signal regardless to the dictionary. For this purpose a phoneme generative automaton was constructed which can synthesize all possible continuous speech model signals for any phoneme sequence. Then the phoneme-by-phoneme recognition of unknown speech signal is applied.

The same context-free phonemes as in baseline recognition system are used.

The experimental accuracy of finding phoneme at the right place equals to approximately 85%.

3.3 Sub-vocabulary retrieval procedure

Preliminary transcription dictionary is prepared to build phoneme triples. The index entry key is a phoneme triple, thus, the index consists of M^3 entries where M is the number of phonemes in the system. Each index entry contains the list of transcriptions that include key phoneme triple. Additional memory usage is approximately 50 MB for vocabulary with 1 M words.

Sub-vocabulary retrieval process is illustrated in Figure 2. Phoneme recognizer output is split into overlapping phoneme triples. Resulting phoneme triple becomes the query. Now, in this system the simple query is used where phoneme triple and query are the same. In the future it should be modified to take into account the insertion, deletion and substitution of phoneme sequence by using *Levenshtein* dissimilarity. Thus phonetic sequence produces the query flow for database.

The query answer consists of the list of transcriptions in which the given triple is included. Next queries produce new transcription portions to be copied into the sub-vocabulary for the second pass. The counter for word repetition is supported to make the rank of word.

All transcriptions in resulting sub-vocabulary are arranged according to the word rank (repetition counter). First N transcriptions are copied into a final sub-vocabulary for the second pass recognition. Thus the recognition sub-vocabulary consists of transcriptions of highest ranks but the vocabulary size does not exceed a fixed limit N.

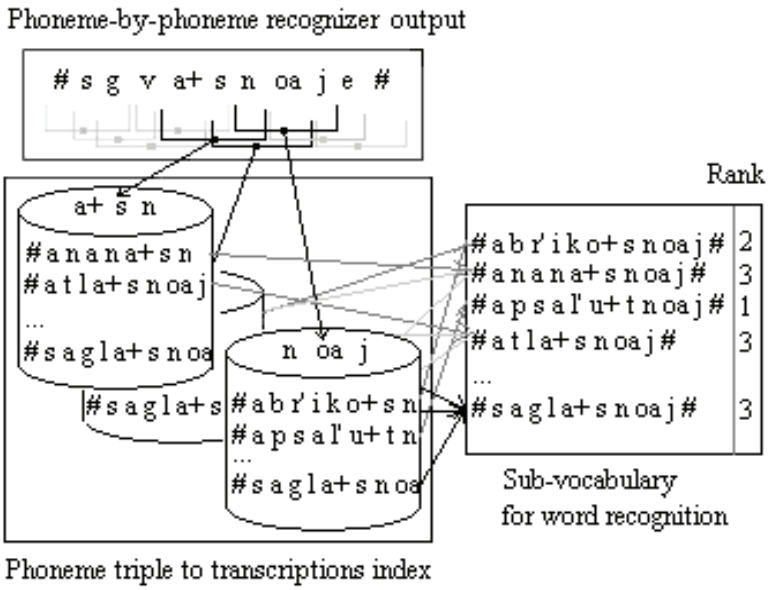

Fig. 2. Sub-vocabulary retrieval process

3.4 ELVIRS algorithm overview

The ELVIRS algorithm (Pylypenko, 2006) works as described in the following.

Preparation stage:

1. Prepare the recognition vocabulary.
2. Chose the phoneme set and build transcriptions for words from vocabulary by rules.
3. Create database index from phoneme triple to transcriptions.
4. Train the acoustic models from collected speech signals.

Recognition stage:

1. Apply phoneme recognizer for input speech signal to produce a phoneme sequence.
2. Split the phoneme sequence into overlapping phoneme triples.
3. Make queries from phoneme triples.
4. Retrieve transcription lists by queries from database index.
5. Arrange transcriptions by the rank.
6. Chose N-best transcriptions for recognition sub-vocabulary.
7. Recognize the input speech signal with sub-vocabulary.

4. The information consideration

The phoneme recognizer output can be considered as a correct phoneme sequence passed through a noisy channel and converted into an output sequence. Denote a right phoneme in output sequence as 1 and wrong one as 0. Let 1 occurs with probability u. The probability P to find k and more successive 1 in a binary set with length of n can be computed with the help of the following recurrent expression:

$$P_n = \begin{cases} 0, n < k \\ u^k, n = k \\ P_{n-1} + u^k(1-u)(1-P_{n-k-1}), n > k \end{cases}$$

Probabilities P to find three and more successive 1 in a binary sequence for different lengths n and probabilities u are shown in Table 3. Average transcriptions length is equal to approximately 8 and the accuracy of finding phoneme at the right place for known utterance is equal to approximately 85%. For these values the probability to find right word in chosen sub-vocabulary is equal 0.953

n \ u	0.75	0.8	0.85	0.9
6	0.738	0.819	0.890	0.948
7	0.799	0.869	0.926	0.967
8	0.849	0.908	**0.953**	0.982
9	0.887	0.937	0.971	0.991
10	0.915	0.956	0.981	0.995

Table 1. Probability to find three and more successive 1 in a binary sequence with length of n

5. ELVIRCOS algorithm for continuous speech

5.1 Architecture

After transcriptions list retrieval procedure an additional procedure – word graph composition is applied. It produces a word network for second pass recognition.

5.2 Word graph composition

The word graph composition procedure is illustrated in Figure 3. Word network starts from vertex S and ends at vertex F. Each triple from phoneme output burns intermediate vertexes with numbers synchronous the occurrence time. On the other hand, each triple became query to data base index, which returns the transcription list as result. Transcriptions are interlaced with intermediate numbered vertexes as base vertexes so that burning phoneme triples are placed in coordination.

The rank of transcription is increased in case when intersection between same transcriptions burned from different phoneme triple occurs. For each moment of time (synchronous with phoneme sequence) the number of involved transcriptions may be calculated.

In order to reduce the word graph complexity, the fixed limit N is applied. For each moment of time transcriptions with small ranks are removed from word graph so that only N transcriptions remain.

The word graph is composed from left to right, that is why it is possible to construct one in real time with the delay is equal of largest transcription length.

5.3 ELVIRCOS algorithm overview

The ELVIRCOS algorithm (Pylypenko, 2007) works as follows.

Preparation stage is the same as ELVIRS algorithm.

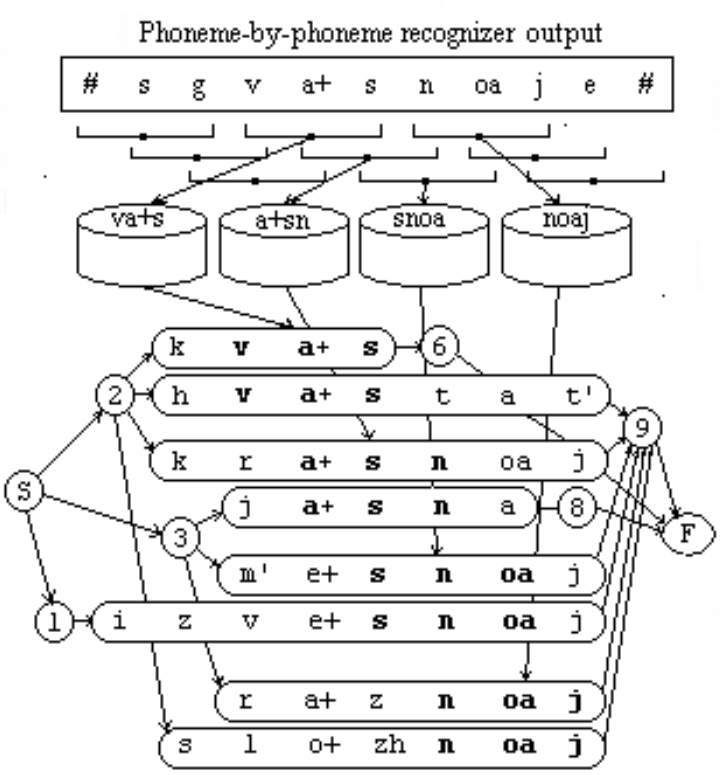

Fig. 3. Word graph composition for continuous speech

Recognition stage:

1. Apply phoneme recognizer to the input speech signal to produce a phoneme sequence.
2. Split the phoneme sequence into overlapping phoneme triples.
3. Make queries from phoneme triples.
4. Retrieve transcription lists by queries from database index.
5. Compose word graph network.
6. Recognize the input speech signal with composed word net.

6. Experimental results

The algorithm was tested at speech corpus from 3 sources:

1. Russian isolated and continuous speech from one speaker with duration 2 hours for training and 20 min for testing.
2. Ukrainian Parliament speech from about 200 speakers with duration 50 hours for training and 3 hours for testing.
3. The November 1992 ARPA Continuous Speech Recognition Wall Street Journal Benchmark Tests.

For experiments, some modifications of HTK or Julius toolkit were necessary to take into account the algorithm.

The considerable reduction of the recognition time (about 10-50 times) with relatively small accuracy degradation (approximately 5%) in comparison with baseline systems has been achieved. The accuracy degradation has a good agreement with the information consideration.
Recognition time not depends from vocabulary size but requires some enlarging because the recognition accuracy fall with vocabulary growth and needs to pay compensation by taking in to account more amount of hypothesis.

7. Future extension

The importance of information retrieval for speech recognition should be underlined. It was shown that additional information source from analysis of phoneme sequence allows to restrict the search space. These new restrictions lead to speech recognition with vocabularies cover practically all words for given language.
Now some modification to adopt bigram language model is developing as a new direction for proposed algorithm. More complex language models can be applied in future works to achieve new features.

8. References

Lee, A., "The Julius book", http://julius.sourceforge.jp, 2009

Pylypenko, V. (2006). "Information Retrieval Based Algorithm for Extra Large Vocabulary Speech Recognition", *Proc. of the 11th International Conference "Speech and Computer", SPECOM'2006*, St. Petersburg, Russia, 2006.

Pylypenko, V. (2007) "Extra Large Vocabulary Continuous Speech Recognition Algorithm based on Information Retrieval", *Proc. of the 8th International Conference "Interspeech 2007"*, Antwerp, Belgium, 2007.

Vintsiuk, T. K. (2000) "Generalized Automatic Phonetic Transcribing of Speech Signals", *Proc. of the 5th All-Ukrainian Conference "Signal/Image Processing and Pattern Recognition"*, pp. 95–98, Kyiv, Ukraine, 2000, in Ukrainian

Vintsiuk, T. K. (2001) "Generative Phoneme-Threephone Model for ASR", *Proc. of the 4th Workshop on Text, Speech, Dialog – TSD'2001*, p. 201, Zelezna Ruda, Czech Republic, 2001.

Young, S.; Kershaw, D.; Odell, J.; Ollason, D.; Valtchev, V. & Woodland, P. (2002). *The HTK Book (for HTK Version 3.2)*. Cambridge University, Cambridge, UK.

Advances in Speech Recognition

Edited by Noam Shabtai

ISBN 978-953-307-097-1
Hard cover, 164 pages
Publisher Sciyo
Published online 16, August, 2010
Published in print edition August, 2010

In the last decade, further applications of speech processing were developed, such as speaker recognition, human-machine interaction, non-English speech recognition, and non-native English speech recognition. This book addresses a few of these applications. Furthermore, major challenges that were typically ignored in previous speech recognition research, such as noise and reverberation, appear repeatedly in recent papers. I would like to sincerely thank the contributing authors, for their effort to bring their insights and perspectives on current open questions in speech recognition research.

How to reference

In order to correctly reference this scholarly work, feel free to copy and paste the following:

Valery Pylypenko (2010). Extra Large Vocabulary Continuous Speech Recognition Algorithm Based on Information Retrieval, Advances in Speech Recognition, Noam Shabtai (Ed.), ISBN: 978-953-307-097-1, InTech, Available from: http://www.intechopen.com/books/advances-in-speech-recognition/extra-large-vocabulary-continuous-speech-recognition-algorithm-based-on-information-retrieval

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821