Information Retrieval Based Algorithm for Extra Large Vocabulary Speech Recognition

Valeriy Pylypenko

Department of Speech and Synthesis
International Research/Training Center for Information Technologies and Systems,
Kyiv, Ukraine
valery_pylypenko@mail.ru

Abstract

This paper presents a new two-pass algorithm for Extra Large (more than 1M words) Vocabulary Speech recognition based on the Information Retrieval (ELVIRS). The principle of this approach is to decompose a recognition process into two passes where the first pass builds the word subset for second pass recognition. With this approach a high performances for large vocabulary speech recognition can be obtained.

1. Introduction

Large vocabulary speech recognition for word-isolated mode has been well investigated at the beginning of 1990s [1]. Different systems for this task have been developed. Further the attention of researchers has concentrated on recognition of continuous speech where due to n-gram word statistics the dictionary volume at each moment of time has been replaced on perplexity - the branching factor for the successive continuation of the text. Such an approach has allowed to distinguish tens thousand words though on each step the choice occurred among hundreds alternatives.

Nevertheless there exists a necessity for speech recognition with a huge numbers of alternatives, in particular for isolated pronounced words.

For example, during the voice control of a computer it is impossible to predict the subsequent word on the basis of several previous because it is defined by control logic, instead of text properties. From other hand there is a necessity for growth the volume of the dictionary to capture all possible speech synonyms of the same command caused by difficulty for users to remember the single command name variant.

A next example concerns the text dictation. Using such systems is limited by the texts, which are statistically similar to one where the n-gram statistics were collected. Additional editing of the collected text demands the presence of all words in the actual dictionary.

Thus, there are many applications where it is desirable to have as large dictionary as possible, in future to cover all words for the given language.

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. The proposed approach has allowed to formulate new two-pass algorithm and to realize it by the program. The fulfilled experiments have shown efficiency of the offered techniques.

2. The baseline recognition system

Proposed method is applicable for any recognition system where phonemes and phoneme-by-phoneme recognition (phonetic transcriber) are present. As a reference system HMM-based HTK toolkit [2] is used.

2.1. Feature Extraction

The speech signal is converted into a sequence of vector parameters with a fixed 25 ms frame and a frame rate of 10 ms. Then each parameter is pre-emphasized with filter $P(z) = 1 - 0.97z^{-1}$. Hamming window is applied. A fast Fourier transform is used to convert time domain frames into frequency domain spectra. These spectra are averaged into 26 triangular bins arranged at equal mel-frequency intervals. 12 dimensional mel-frequency cepstral coefficients (MFCCs) are obtained from cosine transformation and lifter. The log energy is also added as the 13^{th} front-end parameter.

These 13 front-end parameters are expanded to 39 front-end parameters by appending first and second order differences of the static coefficients.

Cepstral mean normalization was applied to deal with the constant channel assumption.

2.2. Acoustic Model

We used Hidden Markov Models (HMMs) with 64 mixtures Gaussian probability density function for acoustic modeling. Acoustic models capture the characteristics of the basic recognition units. All units are modeled by 3 left-to-right states with skip transition. 47 Russian context-free phonemes with pause unit are chosen.

The pronunciation dictionary was created automatically from word orthography using a set of Russian context sensitive rewrite rules.

The Viterbi algorithm is used for the pattern matching.

2.3. Baseline performance

All acoustic models were trained by using 10000 utterances from 2000 words dictionary. All utterances were collected by the single speaker. Isolated word recognition for 1000 utterances from the same speaker is carried out on P-IV 2.4GHz.

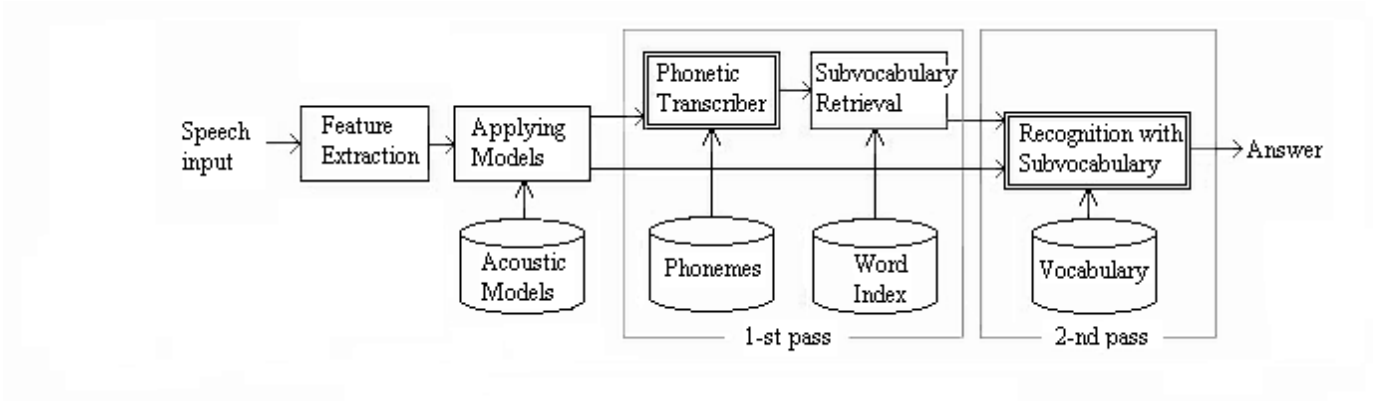

Figure 1: *The architecture of ELVIRS recognition system*

The word recognition accuracy and average time per one second of speech for different vocabulary sizes are given in table 1. Recognition time is linear from vocabulary size. It may be evaluated as 2300 sec for 1987K vocabulary.

Table 1: *Baseline performance*

Vocabulary size	1K	15K	95K
Accuracy, %	99.9	97.9	94.7
Time, sec	1	16	115

3. Proposed Algorithm

3.1. Architecture

The architecture of the system is shown in figure 1. Building blocks like *features extraction*, *applying models* reused from baseline system. *Pattern matching* with subset of vocabulary is used on the second pass also. 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 second pass.

3.2. Phonetic transcriber

The phonetic transcribing algorithm [3, 4, 5] builds a phonetic sequence for speech signal regardless to the dictionary. For this purpose it was constructed some phoneme generative automata 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 accuracy of finding phoneme at the right place for known utterance equals to approximately 85%.

3.3. Sub-vocabulary retrieval

Preliminary transcription dictionary is prepared to build the phoneme triple - transcription index. The index entry key is 3 successive phonemes. So index consists of M^3 entries where M is the number of phonemes in the system. Each entry contains the list of transcriptions that include key phoneme triple.

Sub-vocabulary retrieval process is illustrated in figure 2.

Phonetic transcriber output is split into overlapping phoneme triples. Resulting phoneme triple becomes the query. Now, in our 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 *Levensteine dissimilarity*. Thus phonetic sequence produces the query flow for database.

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

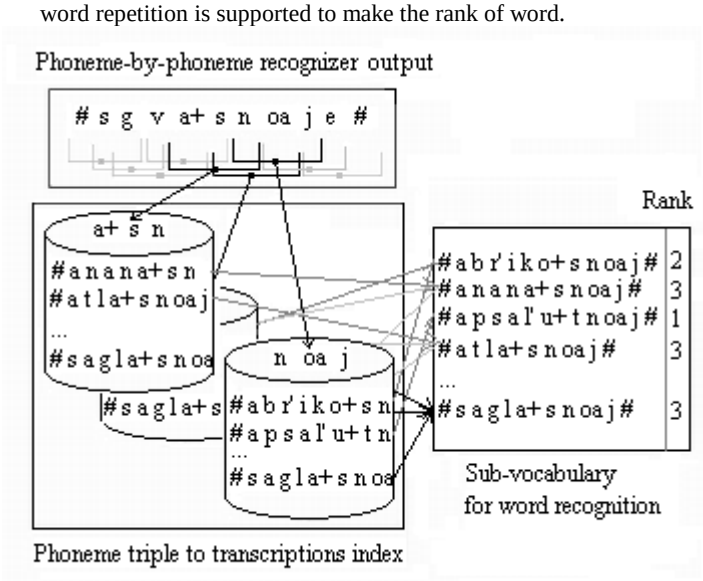

Figure 2: *Subvocabulary retrieval process*

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. So the recognition sub-vocabulary consists of highest ranks of transcriptions but its size does not exceed a fixed limit N.

3.4. Algorithm ELVIRS

The ELVIRS algorithm looks like follows.
Preparing stage:

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

Recognition stage:

- Apply phoneme-to-phoneme recognizer (phonetic transcriber) for input speech signal to produce the phoneme sequence.
- Split phoneme sequence into overlapping phoneme triples.
- Make queries from phoneme triples.
- Retrieve the transcriptions lists by queries from database index.
- Arrange transcriptions by the rank.
- Chose first highest ranked transcriptions with fixed limit N as recognition sub-vocabulary.
- Recognize the input speech signal with sub-vocabulary.

4. Experimental results

Some modifications of HTK toolkit are made to take into account the first pass of algorithm.

The influence of the sub-vocabulary limit N on the average recognition time and word recognition accuracy for different vocabulary sizes was investigated.

Table 1: *ELVIRS recognition performance*

Vocabulary size / Sub-vocab. limit N	15K		95K		1987K	
	Acc, %	Time, sec.	Acc, %	Time, sec	Acc, %	Time, sec
50	92.2	1.4	81.0	1.4	69.2	1.6
100	93.6	1.5	83.9	1.5	72.1	1.7
200	94.6	1.6	87.6	2.1	76.0	1.9
500	**95.5**	**1.9**	**90.1**	**2.5**	80.0	3.3
1000	96.0	2.1	90.7	3.1	82.7	4.4
2000	96.0	4.4	92.0	4.5	**84.8**	**6.8**
5000	96.0	4.6	92.9	8.3	86.4	12.0

The considerable reduction of the recognition time with relatively small accuracy degradation (approximately 5%) in comparison with baseline system has been achieved.

5. Conclusions

In this paper a new approach for large vocabulary speech recognition is described. 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 allow to recognize vocabulary which cover practically all words for given language.

Real time implementation requires context-independent acoustic models. The second pass answer delay can be evaluated as approximately equal the word maximum pronunciation time.

Future work includes the investigation of the influence of phoneme appearance order on the word rank. Some modifications of proposed algorithms for continuous speech will be introduced in the future.

6. References

[1] L. Bahl, P. Brown, P. De Souza, R. Mercer, and M. Picheny, "Acoustic Markov models used in the tangora speech recognition system," *in Proc. ICASSP'88*, NewYork, NY, 1988.

[2] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, V. Valtchev and P. Woodland, "The HTK Book", *Cambridge University Engineering Department*, 2002.

[3] Taras K. Vintsiuk. "Generative Phoneme-Threephone Model for ASR", *Proc. of the 4th Workshop on Text, Speech, Dialog—TSD'2001*, Zelezna Ruda, Czech Republic, 2001, p 201.

[4] Taras K. Vintsiuk. "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.

[5] Valeriy Pylypenko. "Applying Phonetic Transcriber for Large Vocabulary Speech Recognition", *Proc. of the 12th International Conference "Automatic 2005", Vol 3, p. 78*, Kharkov, Ukraine, 2005, in Ukrainian.