Language model comparison for Ukrainian real-time speech recognition system

No Author Given

No Institute Given

Abstract. This paper describes a real-time speech recognition system for Ukrainian designed basically for text dictation purpose targeting moderate computation requirements. The research is focused on language model parameter estimation. As a Slavonic language Ukrainian is highly inflective and tolerates relatively free word order. These features motivates transition from word- to class-based statistical language model. According to our experimental research, class-based LMs occupies less space and potentially outperform a 3-gram word-based model. We also describe several tools developed to visualize HMMs, to predict word stress, and to manage cluster-based language modeling.

1 Introduction

Specific features of Slavonic languages are high inflectiveness and relatively free word order, which leads to rapid growth of the recognition vocabulary (6-8 times larger for same domain in English) and weakening of the language model prediction force. That is why the applicability of conventional methods and algorithms to Slavonic languages looks rather unpromising that is the reason of search for alternative to conventional recognition schemes, particularly considering word composition by the acoustic phoneme decoding output [1]. However, the potential of the recognition scheme having been developed for decades still remains uncovered [2].

The open question is limits of the vocabulary used in the speech-to-text system based on the conventional recognition scheme provided that the system shows real-time performance on a computational platform available for an ordinary user.

Therefore we aimed to build a real-time system that could be exploited on a contemporary personal computer for speech-to-text conversion like a dictation machine.

The system operating conditions must meet potential user's expectations. The recognition vocabulary should cover arbitrary speech with OOV < 1% and means to update the vocabulary must be provided. Acoustically, the system must be able to process speech of every adequate user. In advance prepared speech, read text and spontaneous utterances should be recognized on a similar level of accuracy. The system must provide an ability for the user to dictate in conditions of home and office inside and perhaps outside.

In previous work [3] we described a speech-to-text system that operated in real time with a 100k vocabulary tightly covering common and news domains (politics, economics, culture, education, sports, and weather). Nevertheless we must work with a vocabulary for million words to reach the desired OOV for the arbitrary speech.

In this paper we explain assumptions concerning language distinctions on acoustical, phonetic and lexical levels, try to clear a prospective to attain the necessary vocabulary size, describe respective developed tools and discuss experimental results.

2 Speech-to-text system structure

The basic speech-to-text conversion system structure is shown in Fig. 1. The real-time component implements *Decoder* that refers to *Data and Knowledge Base* developed off-line by means beside the illustrated components.

To create a speech recognition system we developed several data and program resources and used the toolkits available on Internet.

Real time component takes the *Input speech signal* from an available source (microphone, network or file system). *Voice activity detector* suggests beginnings of speech segments for *Pre-processor* that extracts acoustic features from. The system uses mel-frequency cepstral coefficients with subtracted mean and accomplished with energy and dynamic components (delta and delta-delta coefficients). *Decoder* compares an input segment with model signal hypotheses, being generated in accordance to acoustic and language models, using a conservative strategy of non-perspective hypotheses rejection [4]. The output, presented as a confusion network, is passed to *Decision Maker* that forms a *Recognition response* considering the history and performing necessary mappings to symbols and actions.

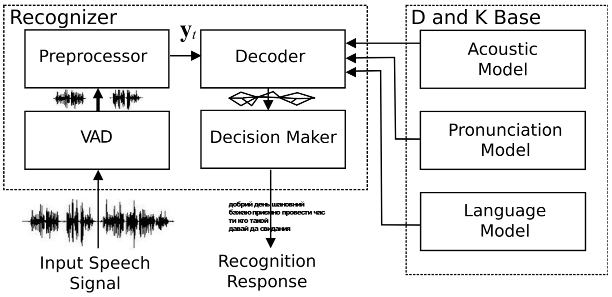

Figure 1: General structure for the basic speech-to-text system.

Acoustic model is developed on a 40 hour subset of the AKUEM speech corpus [5],[6]. The basic phoneme alphabet consists of 56 phonemes including stressed and unstressed versions for 6 vowels. The reason we distinguish them is discussed in next chapter. Currently, HMMs built for context-independent phonemes contain from 8 to 32 Gaussians.

Pronunciation model provides *Decoder* with word pronunciation transcriptions formed off-line by Grapheme-to-phoneme module that implements a multilevel multi-decision symbol conversion technique based on describing the regularities of relation between orthographic and phonemic symbols [7]. An expert formulates about 40 local rules of grapheme-to-phoneme mapping partially modeling the individual speaker peculiarities and co-articulation and reduction of sounds in a speech flow. The rules are adjusted so that on average each word produces about 1.2 transcriptions. The same algorithm with another rules allows for converting numbers, abbreviations and symbolic characters to word sequences. The vocabulary for the entire system consists of a frequency dictionary extracted from the large text corpus and supplementary vocabularies covering speech corpus, social and local dialects, proper names, abbreviations etc. Taking a specified amount of top-frequent words from the system vocabulary a recognition vocabulary is formed.

Language model is created proceeding from the recognition vocabulary and a text corpus subset consisting of sentences containing below the specified portion of OOV words. The basic text corpus is derived from a hypertext data downloaded from several websites containing samples of news and publicity (60%), literature (8%), encyclopedic articles (24%), and legal and forensic domain (8%). To be noted that the data downloaded from news websites contains numerous user comments and reviews, which we consider as text samples of spontaneous speech. Text filter, used for text corpus processing, provides conversion of numbers and symbolic characters to relevant letters, removing improper text segments and paragraph repetitions. Total size of the basic text corpus is 2 GB that includes 17.5 million sentences that is a list of words containing above 275 million items and forming a vocabulary of more than two million words.

For the recognition vocabulary of 100 000 words, 88.5 million distinct 3-grams are detected in the subset of the basic text corpus after removing sentences containing more than 20% or at least three running unknown words. This sub-corpus is used for language modeling and referred as 250 M corpus. Consequently, we got OOV words occupy 2.5% of all words that is about twice less than in Ukrainian arbitrary text for the specified vocabulary size. To model spontaneous speech characteristics a class of transparent words is introduced to the recognition vocabulary. It contains non-lexical items like pause fillers and emotion and attitude expressions (laugh, applauds etc.).

Applying language modeling tool [8] we have received a text file in ARPA format that occupies 5 GB reduced to 1.2 GB by a module of the decoder tool [4].

The real-time modules are used to build a basic speech-to-text conversion system for experimental research and trial operation. Graphical user interface integrated with the basic system allows for demonstrating continuous speech recognition for wide domain in real time, using a contemporary notebook [3].

Further, we consider a transition from word- to class-based statistical language model in order to move towards a vocabulary that provides the desired OOV for the arbitrary speech.

3 Class-based LM development

As a Slavonic language, Ukrainian is highly inflective, the number of word forms per dictionary entry accedes 12 that is about 6 times more than for English. Therefore, to build an adequate language model a 6 time larger vocabulary is required. Moreover, relatively free word order is normative that leads to perplexity and data sparsity growth. Analysis of these features motivates a transition from word- to class-based statistical language model that operates with transition probability and membership probability [9].

Word clustering procedure tries to maximize the perplexity improvement criterion

$$F_{\boldsymbol{G}} = \sum_{g,h\in G} C(g,h)\log C(g,h) - 2\sum_{g\in G} C(g)\log C(g) \tag{1}$$

where (g, h) means a class g follows a class h from the set of equivalence classes $\boldsymbol{G}$ and function $C(\cdot)$ counts its argument occurrence in the training corpus. An exchange algorithm described in [9] implies iterations in which each word is tested for a better class and consequently moved there. While implementing the algorithm we came to an alternative formulation of criteria computation refinement.

Let us enumerate all equivalence classes: $g_i \in \boldsymbol{G} \quad i=1:G$ and introduce $C_{ij}=C(g_i, g_j)$ for successor and C^{-}_{ij} for predecessor occurrence.

Assuming that a preceding single classification function $G^{-}(\cdot)$ applied to w has given g_u, i.e. $G^{-}(w)=g_u$, we are to check a hypothesis of transition w to another class indexed with v, i.e., $G(w)=g_v$.

The first sum in (1), having the most complicated computations, $O(G^2)$, can be expressed as

$$\sum_{i,j}\log C_{ij} = \sum_{\substack{i,j\\ (i,j)\cap(u,v)=\emptyset}} \log C_{ij} + \sum_{\substack{i=u,v\\ j}} \log C_{ij} + \sum_{\substack{j=u,v\\ i\neq u,v}} \log C_{ij}. \tag{2}$$

Thus, the analyzed sum is decomposed in three components where the most expensive for computations component, still $O(G^2)$, might be expressed as a recursion relatively to the predecessor:

$$\begin{aligned} \sum_{\substack{i,j\\ (i,j)\cap(u,v)=\emptyset}} \log C_{ij} &= \sum_{\substack{i,j\\ (i,j)\cap(u,v)=\emptyset}} \log C^{-}_{ij} = \\ &= \sum_{i,j}\log C^{-}_{ij} - \left(\sum_{\substack{i=u,v\\ j}} \log C^{-}_{ij} + \sum_{\substack{j=u,v\\ i\neq u,v}} \log C^{-}_{ij} \right) \end{aligned} \tag{3}$$

arriving to the computation time complexity of $O(G)$. Proceeding from (1)–(3) we have developed an efficient tool for word clustering and assigning a new word, accomplished with bigram counts, to one of existing classes.

The clustering results have been analyzed proceeding from their relevance to linguistic categories. Firstly automatically obtained classes for Ukrainian in general correspond to syntactic, semantic and phonetic features.

Most word classes have an obvious syntactic interpretation, such as nouns in a genitive form, or plural adjectives. Table 1 shows several word classes that have been obtained by bigram clustering on the 250 M corpus for 1000 word classes. The words in each word class are listed in descending word unigram count order and the most frequent word is emphasized. We present three classes completely and first 7 words for the last class.

Table 1. Bigram clustering examples, G = 1000

Words of cluster with meaning	Frequency
багато / many, much	134590
чимало / plenty	24482
безліч / a lot of	7696
немало / quite a lot of	2191
якнайбільше / as many	760
багацько / lots of	255
богато (*misspelled* багато)	123
які / that, which (plural)	590681
котрі / that, which (plural)	24499
яки (*misspelled* які)	465
де / where	246376
куди / to where	31966
звідки / where from	15373
звідкіль / where from (colloquial)	120
заявив / [he] stated	163547
вважає / [he, she] supposes	99803
повідомив / [he] informed	80043
заявила / [she] stated	32795
заявляє / [he, she] states	31965
розповів / [he] told	30504
говорить / [he, she] speaks	29756

Often, there is some semantic meaning like in the last class containing verbs of communication (for third person in present and past tenses). Two first classes show that misspelled but still frequent words may join to the class containing a correct version of the word.

In Ukrainian, words may have different forms in dependence of phonetic context. For instance, the conjunction *and* has three forms normally used between consonants, between vowels and in other cases. All these forms were automatically assigned to different classes.

4 Data

The basic dictionary is extracted from the electronic lexicography system subset containing 151 962 lemmas, including over 10 thousand names, that makes 1.90 million word forms [10]. Due to shared spelling the actual word form vocabulary consists of 1.83 million words that have different either spelling or primer lexical stress position.

The basic text corpus is derived from a hypertext data downloaded from several websites containing samples of news and publicity (60%), literature (8%), encyclopedic articles (24%), and legal and forensic domain (8%). To be noted that the data downloaded from news websites contains numerous user comments and reviews, which we consider as text samples of spontaneous speech. Text filter, used for text corpus processing, provides conversion of numbers and symbolic characters to relevant letters, removing improper text segments and paragraph repetitions. Hereafter, we refer to the basic corpus as 275M corpus. In accordance to the corpus summary shown in Table 1, we observe 6.64 word forms per lemma in average, whereas this relation is twice greater, 12.3, within the dictionary [10]. Adding 200 000 most frequent words to the vocabulary we reduce OOV to less than 0.5%.

Table 2. Basic text corpus 275M summary

Words	Sentences	Vocabulary			OOV	Homographs
		All words	Known words	Known lemmas		
275 288 408	1 752 371	1 996 897	801 040	120 554	2,51%	16 729 476

Words that have 2 or more valid stress positions, referred as homographs, take over 6% of the average text. While estimating acoustic model parameters all stress versions of homographs were used on realignment stage.

5 Experiments

We evaluated three language model types on two, however, relatively small test sets with different OOV. Error rates considered are based on both words (WER) and characters (CER). Vocabulary size was set to 100 000 words, in average, 1.1 pronunciations per word were generated. Word-based 3-gram language model is denoted as w3 and class-based language models, c3 and c4, are built respectively for 3- and 4-grams. Size of 3-gram class-based LM is 9 times less than word-based LM, 4-gram class-based LM occupies somewhat less space than 3-gram word-based model.

According to our experimental research shown in Table 3, class-based LMs have a certain potency due to character error close to or even smaller than the word-based LM error rate. Phonetically close words assigned to same class is a source of mistakes, as far the word with much better membership probability may get better chance to win. This could be compensated by stimulating such words to stay in different classes.

Table 3. Speech recognition summary for different language models

LM	Error type	OOV = 3.4	OOV = 2.6
w3	WER	20.9	15.9
	CER	7.8	4.2
c3	WER	28.2	24.3
	CER	6.9	6.0
c4	WER	28.6	24.6
	CER	10.0	6.5

6 Conclusion

The described real-time system for Ukrainian speech-to-text conversion demonstrates a potential of focusing on language distinctive features, which makes feasible to attain vocabulary size necessary to reduce OOV below 1% and to introduce punctuation and character case dependency.

For dictating purpose, human-machine interaction is crucial. The system has to suggest recognized utterance refinement based on multi-decision recognition response; moreover, accepted refinements must update the recognition response model. Besides assigning a new word to the unknown word category, we plan to implement updating the class language model by mapping new words to classes and recomputing membership probabilities.

Distance to closest extrinsic classes should give a clue to predicting homographs and consequent semantic word decomposition that may lead to more homogeneous classes.

For text processing, more precise number and symbol to grapheme conversion is topical in order to predict their correct concordance for the observed context.

The development of the presented system is on early stage. In near future several improvements will be completed, which will increase accuracy and extend the scope of usage.

References

1. Taras Vintsiuk, Mykola Sazhok. Multi-Level Multi-Decision Models for ASR. In Proc. SpeCom'2005, Patras, 2005, pp.69-76.
2. M. Gales and S. Young. "The Application of Hidden Markov Models in Speech Recognition." Foundations and Trends in Signal Processing, 2007, 1(3), pp. 195- 304.
3. Blind review.
4. A. Lee, T. Kawahara. Recent Development of Open-Source Speech Recognition Engine Julius. APSIPA ASC, 2009, pp. 131-137.
5. Young S.J. et al., The HTK Book Version 3.4, Cambridge University, 2006.
6. Valeriy Pylypenko, Valentyna Robeiko, Mykola Sazhok, Nina Vasylieva, Oleksandr Radoutsky. Ukrainian Broadcast Speech Corpus Development // Specom'2011, Kazan'2011, pp. 244-247.
7. Blind review.
8. Bo-June (Paul) Hsu and James Glass. Iterative Language Model Estimation: Efficient Data Structure & Algorithms. In Proc. Interspeech, 2008.
9. S. Martin, J. Liermann, and H. Ney, "Algorithms for bigram and trigram word clustering," in Proceedings of Eurospeech, vol. 2, pp. 1253–1256, Madrid, 1995
10. http://lcorp.ulif.org.ua/dictua/